LaVPR: Benchmarking Language and Vision for Place Recognition

Ofer Idan, Dan Badur, Yosi Keller, and Yoli Shavit*

Faculty of Engineering, Bar-Ilan University
Ramat Gan, Israel
{ofer.idan, badur.dan, yosi.keller, yoli.shavit}@biu.ac.il

Abstract. Visual Place Recognition (VPR) often fails under extreme environmental changes and perceptual aliasing. Beyond these limitations, standard systems cannot perform 'blind' localization from verbal descriptions alone, a capability critical for applications such as emergency response. To address these challenges, we introduce LaVPR, a large-scale benchmark that extends existing VPR datasets with over 650,000 rich natural-language descriptions. Using LaVPR, we investigate two paradigms: Multi-Modal Fusion for enhanced robustness and Cross-Modal Retrieval for language-based localization. Our results show that language descriptions yield consistent gains in visually degraded conditions, with the most significant impact on smaller backbones. Notably, adding language allows compact models to rival the performance of much larger vision-only architectures. For cross-modal retrieval, we establish a baseline using Low-Rank Adaptation (LoRA) and Multi-Similarity loss, which substantially outperforms standard contrastive methods across vision-language models. Ultimately, LaVPR enables a new class of localization systems that are both resilient to real-world stochasticity and practical for resource-constrained deployment. Our dataset and code are available at https://github.com/oferidan1/LaVPR.

Keywords: Language-Vision Place Recognition · Multi-Modal Fusion · Cross-Modal Retrieval

1 Introduction

Traditional Visual Place Recognition (VPR) relies on matching global image descriptors to retrieve geo-tagged database entries. However, even state-of-the-art visual models struggle under extreme environmental variations, such as adverse weather, long-term temporal shifts, and motion blur [26]. To achieve robust, interactive localization, we propose extending the VPR framework with natural language, focusing on two key paradigms:

i) *Multi-Modal Fusion (V + L)*: This approach utilizes language as a stable, high-level descriptor to disambiguate visual scenes. By incorporating semantic

* Corresponding Author.

scene elements and visible text directly into a descriptive narrative, such as *"a brick building featuring a 'Pharmacy' sign above the entrance"*, the system gains a viewpoint-invariant context. While visual degradation can compromise pixel-level features, these linguistic anchors provide a stable reference that maintains retrieval accuracy where vision-only models fail.

ii) *Cross-Modal Retrieval ($L \rightarrow V$):* This paradigm enables "blind" localization, where a system recognizes a place from a textual description alone. This capability is critical for semantic robotics (interpreting human instructions), forensic geolocation (witness descriptions), and emergency rescue calls. For instance, a distressed caller might describe high-stress conditions, e.g., *"someone fainted and I am near a tall stone building with a clock tower"*, requiring the system to geolocate the scene solely through linguistic cues. This task requires learning a shared latent space that can align abstract verbal cues with the specific, fine-grained visual features necessary for precise localization.

While these paradigms offer a path toward more resilient localization, mainstream VPR benchmarks are image-only, lacking the textual descriptions required to train and evaluate vision-language models. To address this, we introduce *LaVPR*, a large-scale benchmark comprising over 650,000 descriptions. LaVPR systematically extends established VPR datasets with linguistic data generated via a Visual Large Language Model (VLLM). To ensure spatial fidelity and reduce hallucinations, we employ a segmentation-grounded pipeline and a human-in-the-loop protocol (Section 3).

Using LaVPR, we provide a comprehensive evaluation of vision-language strategies. For Multi-Modal Fusion (Section 4), our results reveal consistent performance gains across a diverse range of visual backbones. Moreover, we find that incorporating linguistic context allows compact visual backbones (e.g., ViT-S) to rival the accuracy of much larger, compute-intensive models. This offers a path toward high-throughput, efficient deployment. For Cross-Modal Retrieval (Section 5), we observe that standard zero-shot transfer and contrastive fine-tuning fail to achieve precise localization. We instead propose a parameter-efficient alignment strategy using Low-Rank Adaptation (LoRA) [14] in conjunction with the Multi-Similarity (MS) loss [39], establishing a robust baseline for cross-modal VPR.

In summary, our contributions are as follows:

- We introduce LaVPR, a standardized, open-source benchmark for vision-language place recognition that extends established VPR datasets with over 650,000 aligned, grounded textual descriptions.
- We demonstrate that language descriptions provide consistent accuracy gains across diverse architectures, showing that linguistic context allows compact backbones to rival the performance of significantly larger vision-only models.
- We present an empirical study on cross-modal alignment for "blind" localization, demonstrating that pairing LoRA with the MS loss is critical for learning a discriminative shared embedding space, substantially outperforming standard contrastive baselines.

2 Related Work

2.1 Visual Place Recognition

VPR systems are defined by three primary design choices: the architectural backbone, the aggregation strategy, and the training objective. Backbones have transitioned from standard CNNs [4, 31] to Transformer-based models that leverage self-attention for superior global context [26]. To compress these features into a global descriptor, methods utilize aggregation strategies such as NetVLAD's residual clustering [4] or Generalized Mean (GeM) pooling [31]. Recent gains involve enhancing these descriptors through attention mechanisms [16], multi-scale aggregation [3, 11], and adapter-based fine-tuning of foundation models [27]. This shift reflects a trend where pre-trained, large-scale encoders provide robust features that often outperform specialized, task-specific architectures. Training typically utilizes contrastive objectives, such as triplet [12] or multi-similarity loss [39], to enforce invariance to environmental changes while maximizing spatial separation. To allow training on dense city-scale datasets, recent methods adopt classification formulations [5, 6] utilizing large-margin cosine losses [37].

The field relies on retrieval metrics (Recall@K) across benchmarks characterized by viewpoint shifts and extreme temporal variations [34, 35, 45]. Large-scale training sets like GSV-Cities [2] and SF-XL [5] have become standard for these tasks. LaVPR builds on these foundations by augmenting these widely-used datasets with over 650,000 aligned textual descriptions. Unlike previous works limited to uni-modal evaluation, this extension enables a controlled benchmark of multi-modal and cross-modal strategies using identical visual backbones, isolating the impact of linguistic context.

2.2 Language-Vision Place Recognition

The exploitation of complementary sensor inputs, and language in particular, remains underexplored for VPR research [24].

Multi-Modal Place Recognition. Prior fusion methods have predominantly integrated 3D geometry (depth or LiDAR) with imagery [15, 20, 21, 23, 30, 46], consistently improving robustness over vision-only approaches. Within the text domain, methods like TextPlace [13] have leveraged scene text through Optical Character Recognition (OCR) as an invariant feature, capitalizing on the distinctiveness of street signs and shop names for location identification.

Most recently, MSSPlace [29] explored fusing diverse data streams, including multi-camera imagery, LiDAR, and text, into a unified global descriptor. To facilitate this, the authors extended the Oxford RobotCar [28] and NCLT [8] datasets using MiniGPT-4 [47]. While this work provided valuable evidence for the benefits of multi-modal fusion, its evaluation protocol lacked a controlled ablation, comparing multi-modal, multi-view inputs against single-view baselines trained on disparate datasets. Furthermore, the absence of a released dataset (to date) or a detailed curation protocol limits the reproducibility of their findings.

Our work addresses these limitations by introducing a rigorous, publicly available framework for standardized language-vision place encoding. By maintaining a strictly controlled experimental setup, we isolate the specific contribution of language, revealing new insights into how semantic descriptions complement visual features to achieve robust VPR.

Cross-Modal Place Recognition. Research in this area has primarily focused on cross-modal text-to-point-cloud localization [25, 41, 44]. Introduced by Kolmet et al. [19], this task aims to regress a 6-DoF pose within a 3D environment using only natural language descriptions. Standard pipelines typically employ a coarse-to-fine approach: mapping textual queries and 3D sub-maps into a shared embedding space for retrieval, followed by fine-grained matching of semantic instances. Evaluation relies on datasets like KITTI360Pose [19], an extension of KITTI360 [43] containing template-based descriptions of spatial relations. In the text-to-image domain, Shang et al. recently proposed a multi-view registration approach to align grouped textual descriptions with visual scenes [33]. However, this method diverges from the standard VPR paradigm by requiring multi-view textual queries and focusing on set-to-set alignment rather than scalable, single-query global descriptor retrieval on established benchmarks. In LaVPR, we directly extend VPR training and evaluation benchmarks with textual descriptions, curate them and then analyze the contribution in a controlled setting.

3 The LaVPR Benchmark

We introduce LaVPR, a large-scale benchmark designed to bridge visual localization and natural language. By augmenting established VPR datasets with 651,865 aligned descriptions (Table 1), LaVPR enables the systematic evaluation of multi-modal fusion and cross-modal retrieval while preserving original geographic splits for fair comparison.

3.1 LaVPR Generation

We employ Gemini 2.5 Flash [10] to generate rich, high-density narratives that prioritize permanent architectural features and spatial relationships over transient elements like pedestrians.

Unlike sparse keyword-based datasets, LaVPR provides dense semantic descriptions (Figure 1) that capture fine-grained textures and scene text. This descriptive depth is summarized in Table 3. To isolate the impact of language when vision is compromised, we further curated specialized subsets targeting visual degradation and discriminative scene text (detailed in Table 2). Section 1 in our Supplementary (Suppl.) Material provides the generation prompt and the specialized dataset generation protocols, including example images.

Table 1: Overview of the VPR datasets integrated into LaVPR. The combined dataset consists of 651,865 image-description pairs. Dataset characteristics are curated based on [26].

Dataset	Type	Usage	Image Count	
			Database	Queries
GSV-Cities [2]	Urban	Train	529,506	–
Pitts30K-Val [35]	Urban, Panorama	Validation	10,000	7,608
Pitts30K-Test [35]	Urban, Panorama	Test	10,000	6,816
AmsterTime [45]	Very long-term	Test	1,231	1,231
MSLS-val [40]	Urban, Suburban	Test	18,871	740
MSLS-challenge [40]	Long-term	Test	38,770	27,092
Total (LaVPR)			**608,378**	**43,487**

Fig. 1: Example image from the Amstertime dataset. Aligned description in LaVPR: 'Leftmost red brick building with an upper window and a black wrought iron balcony railing, cream-colored framed storefront with a display window, recessed wooden entrance door, narrow brick wall section with house number '3', grey vertical downspout, middle cream-colored framed shop with a glass door, large storefront window displaying "de Witte TandenWinkel", black lower facade panel, red brick upper floor with a rectangular window, blue square sign, grey electrical box, rightmost red brick building with an upper window, yellow scalloped awning labeled "KAAS CHEESE".'

Table 2: Specialized subsets in LaVPR. These datasets target high linguistic discriminativeness and extreme visual domain gaps to isolate the impact of the language channel.

Dataset	Key Query Properties	Reference Database	# Queries
Amstertime-La	Scene text	Amstertime	31
MSLS-Blur	Scene text, blur	MSLS-val	86
MSLS-Weather	Scene text, adverse weather	MSLS-val	86

Table 3: Key linguistic and structural statistics of the LaVPR benchmark. These metrics highlight the descriptive density and lexical diversity across the integrated datasets.

Metric	Training	Validation	Test	Total
Total Image-Text Pairs	529,506	17,608	104,751	**651,865**
Avg. Word Count per Caption	52.78	50.35	46.43	51.69
Avg. Entity Count (Nouns)	23.66	22.49	21.35	23.25
Total Unique Vocabulary	176,893	16,455	55,110	209,200
Signage/Text-heavy Samples (%)	27.77%	55.66%	30.06%	28.89%

3.2 LaVPR Curation

While modern VLMs generate rich descriptors, they remain susceptible to hallucinations [17]. To rigorously assess and ensure data reliability, we developed a comprehensive validation pipeline that integrates automated *entity extraction* (Phi-3.5-mini-instruct[1]), *spatial grounding* (SAM3[2]), and *binary VLM verification* (Qwen2-VL-7B-Instruct[3]). To maximize safety, the automated stage is tuned for exceptionally high recall (91%) over precision (22%), as measured on a manually curated validation set. This stringent tuning ensures that virtually all potential discrepancies are flagged for review, albeit with a high rate of false positives. We resolve these ambiguities through a Human-in-the-Loop (HL) phase, distinguishing genuine hallucinations from false alarms caused by out-of-distribution elements or visual noise.

We deployed this auditing protocol across our evaluation splits to verify benchmark fidelity. Crucially, while the hyper-sensitive automated filter flagged 26% of samples for secondary verification, subsequent HL analysis of the largest unresolved groups revealed that the raw data is inherently of high quality: genuine hallucinations were rare, affecting only $\sim 1\%$ of the inspected images. Typically, these minor cases involved a single erroneous object within otherwise precise descriptions; these rare confirmed hallucinations were subsequently purged from the evaluation sets. Furthermore, manual validation of in-scene text confirmed a 93.2% accuracy rate, with minor errors isolated to severe occlusions.

We emphasize that all main-paper results are obtained using the raw, uncurated training data. The curation pipeline described above was used purely as a diagnostic audit to guarantee evaluation integrity, rather than as a necessary cleaning step for model optimization. Crucially, achieving strong performance on raw, VLM-generated data demonstrates both the high baseline quality of LaVPR and our framework's inherent robustness to minor visual-language noise. For completeness, a comprehensive algorithmic walkthrough of the curation pipeline, step-by-step statistical breakdowns, and an ablation study of our cleaning strategy are provided in Section 2 of the Suppl. Material.

[1] `https://huggingface.co/microsoft/Phi-3.5-mini-instruct`, [1]
[2] `https://huggingface.co/facebook/sam3`, [7]
[3] `https://huggingface.co/Qwen/Qwen2-VL-7B-Instruct`, [38]

4 Multi-Modal Place Recognition ($V + L$)

We investigate the synergy between natural language and vision to address the inherent fragility of image-only descriptors. By integrating linguistic context, we provide a stable, high-level semantic reference that persists across temporal and environmental shifts.

4.1 Late-Fusion Methods

We evaluate four late-fusion architectures utilizing frozen, pretrained visual (E_v) and textual (E_t) encoders. This design isolates cross-modal dynamics and demonstrates the potential of language as a "robustness anchor" without requiring extensive end-to-end retraining.

Given an image I and description t, we obtain embeddings $z_v = E_v(I) \in \mathbb{R}^{d_v}$ and $z_t = E_t(t) \in \mathbb{R}^{d_t}$. We explore the following fusion operations:

- Concatenation (CAT): A baseline where $z_e = [z_v; z_t] \in \mathbb{R}^{d_v+d_t}$.
- Projection and Addition (PA): Linear layers project both modalities to a shared dimension d_e, where $z_e = \text{Linear}(z_v) + \text{Linear}(z_t)$.
- Multi-Layer Perceptron (MLP): A shallow MLP processes the concatenated vector, expanding the dimension to $(d_v + d_t)$ before mapping to a fixed latent space d_e.
- Adaptive Score Fusion (ADS): Instead of feature fusion, we propose to adaptively weigh modality-specific similarity scores (depicted in Fig. 10 in the Suppl. Material). An MLP with Softmax predicts weights $W = [w_v, w_t]$ from concatenated visual and language query features. For a query-reference pair (i, j), weights are averaged: $w_{m_{ij}} = (w_{m_i} + w_{m_j})/2$ for $m \in \{v, t\}$. The joint similarity is $S_{ij} = w_{v_{ij}} \cdot S_{v_{ij}} + w_{t_{ij}} \cdot S_{t_{ij}}$.

All fusion modules, except for CAT, are optimized using the Multi-Similarity (MS) loss [39]:

$$\mathcal{L}_{MS} = \frac{1}{|B|} \sum_{q \in \mathcal{B}} \Bigg\{ \frac{1}{\alpha} \log \Big[1 + \sum_{p \in \mathcal{P}_q} e^{-\alpha(S_{qp} - \lambda)} \Big] + \frac{1}{\beta} \log \Big[1 + \sum_{n \in \mathcal{N}_q} e^{\beta(S_{qn} - \lambda)} \Big] \Bigg\}, \tag{1}$$

where for each query q in a batch B, $\mathcal{P}_q$ and $\mathcal{N}_q$ correspond to the sets of positive and negative samples for q, respectively, and S_{qp} and S_{qn} are the cosine similarities of a positive pair $\{q, p\}$ and a negative pair $\{q, n\}$, correspondingly. α and β and λ are hyper-parameters. In ADS, we use the multi-modal similarity scores as input to the MS loss.

To distill salient linguistic features, we further introduce a Learned Language Pooling (LLP) module. This module employs a self-attention mechanism over the hidden states of E_t to generate a context-aware textual representation. We evaluate this pooling mechanism in combination with the four late fusion strategies. Additional implementation details are provided in Section 5 in the Suppl. Material.

4.2 Results & Discussion

We evaluate multi-modal fusion across diverse VPR models, including NetVLAD [4], CosPlace [5], EigenPlaces [6], MixVPR [3], SALAD [18], and CricaVPR [26], using BAAI/bge-large-en-v1.5 (BGE-L) model [42] as the primary text encoder.

Evaluation of Fusion Mechanisms: As shown in Table 4, ADS and CAT emerge as the most effective multi-modal integration strategies when paired with MixVPR. While CAT achieves competitive performance across datasets, ADS yields peak robustness when paired with LLP. For example, ADS with LLP achieves the highest absolute performance on Amstertime, improving R@1 to 38.1%. Conversely, CAT performance degrades when combined with LLP, indicating that fixed structural concatenation cannot accommodate independently projected modalities without dynamic rescaling, which ADS naturally handles via adaptive scoring. We further show that CAT and ADS provide consistent gains across different textual encoders in our Suppl. Material (Section 4, Table 4).

Table 4: Comparison of late-fusion mechanisms using MixVPR (embedding dimension of 512) and BGE-L as the visual and textual backbones, respectively. Shading indicates the best and runner-up multi-modal configurations per dataset.

Fusion	**Amstertime**			**MSLS-val**			**Pitts30**		
Mechanism	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Visual-only	35.7	53.1	60.4	83.2	91.9	93.4	90.6	95.6	96.3
Without LLP									
ADS	36.1	53.4	60.5	**83.4**	**91.9**	93.4	90.7	**95.6**	96.3
CAT	37.0	58.2	**65.6**	81.9	91.5	93.0	89.9	95.3	**96.4**
MLP	31.7	52.6	59.8	83.2	91.5	93.0	89.8	95.1	96.2
PA	31.0	50.8	57.3	83.0	90.1	91.9	89.7	95.1	96.3
With LLP									
ADS	**38.1**	**58.4**	64.7	83.2	91.6	**93.5**	**90.8**	95.4	**96.4**
CAT	30.2	51.0	59.7	77.4	89.1	91.6	85.7	93.9	95.8
MLP	32.6	50.8	58.3	**83.4**	90.5	92.8	89.6	95.1	96.2
PA	30.2	47.5	54.9	82.3	90.1	92.3	89.6	94.8	96.1

Influence of Pre-training Paradigms and Backbone Architecture: Beyond the fusion mechanism itself, the magnitude of linguistic gains is heavily governed by the architecture of the visual backbone and its pre-training objective (Table 5). Supervised convolutional backbones (VGG and ResNet) exhibit the most dramatic relative improvements, with NetVLAD achieving up to a 200% R@1 gain, indicating that classification-based pre-training leaves substantial "semantic gaps" regarding environmental context that text effectively bridges.

Conversely, self-supervised (SSL) Vision Transformer backbones like DINOv2 capture a much richer visual-semantic prior, naturally narrowing the margin

Table 5: Relative R@1 Gain (%) of Language-Augmented (La-) methods compared to visual-only baselines. Positive gains indicate linguistic complementarity; negative values indicate slight degradation. Embedding dimension for each backbone is denoted in brackets (d). VGG16 and ResNet50 backbones are pretrained on ImageNet-1k in a supervised manner. DinoV2-L is pretrained via invariance-based self supervised learning (SSL). VPR Loss Abbreviations: TRP: Triplet Loss; LMC: Large Margin Cosine Loss; MS: Multi-Similarity Loss.

Method	Fusion	Backbone (d)	Loss	Relative R@1 Gain (%) per Dataset				
				MSLS-B	MSLS-W	Ams.	MSLS-C	Pitts30
La-NetVLAD	CAT	VGG16 (32k)	TRP	+162.3	+136.3	+199.5	+18.3	-1.0
La-CosPlace	CAT	ResNet50 (512)	LMC	+44.9	+29.4	+13.2	+5.7	+0.3
La-EigenPlace	CAT	ResNet50 (2k)	LMC	+43.2	+15.3	+15.5	+3.4	+0.2
La-MixVPR	CAT	ResNet50 (4k)	MS	+21.3	+14.9	+6.0	+0.9	-0.9
La-MixVPR	ADS+LLP	ResNet50 (4k)	MS	+14.8	+10.4	+11.9	+2.2	+0.3
La-SALAD	CAT	DinoV2-L (8k)	MS	+1.2	+3.9	+5.5	-0.9	-0.8
La-SALAD	ADS+LLP	DinoV2-L (8k)	MS	+1.2	+3.9	+5.5	0.0	+0.1
La-CricaVPR	CAT	DinoV2-L (10k)	MS	0.0	+5.2	-9.9	-3.7	-1.5
La-CricaVPR	ADS+LLP	DinoV2-L (10k)	MS	+1.2	+6.6	0.0	+0.6	+0.1

for linguistic improvement ($< 7\%$). Yet, even within this saturated regime, our proposed ADS+LLP fusion framework yields consistent, non-trivial advantages where simple CAT fusion plateaus.

Compensatory Dynamics in Adverse Environments: The utility of language as a compensatory signal becomes most pronounced under adverse conditions where visual reliability is compromised (Table 6). Language-augmented (La-) methods consistently outperform their visual-only baselines for evaluation subsets characterized by extreme blur and severe weather. In these degraded scenarios, the textual description functions as a robust semantic anchor that remains completely invariant to pixel-level visual domain shifts. A qualitative example illustrating how La-MixVPR successfully resolves an environmental degradation that misleads baseline MixVPR is shown in Fig. 2. .

A similar trend is observed also when utilizing the CAT fusion (Suppl. Material, Section 4, Table 3). However, while CAT is competitive for compact descriptors, the ADS-LLP framework provides greater scalability and robustness for high-dimensional backbones like CricaVPR.

Computational Efficiency and Scaling: Language integration further offers a path toward efficient deployment. As shown in Table 7, by pairing a compact visual backbone (CricaVPR with ViT-Small) with a lightweight text encoder (BGE-Small), we achieve 93.0 R@1 on MSLS-Blur, outperforming the massive visual-only CricaVPR-Base. This hybrid approach reduces computational complexity by

Table 6: Comparison of vision-only VPR methods vs. our Language-Augmented (La-) variants using ADS+LLP Fusion and BGE-L. Bold indicates the best performance within each backbone pair.

Method	Dim	MSLS-B			MSLS-W			Amst.-La			MSLS Ch.		
	(*d*)	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
MixVPR	4096	70.9	82.6	87.2	77.9	83.7	88.4	54.8	80.6	83.9	63.4	76.0	79.2
La-MixVPR	+1024	**81.4**	**93.0**	**97.7**	**86.0**	**91.9**	**96.5**	**61.3**	**90.3**	**90.3**	**64.8**	**76.9**	**80.3**
SALAD	8448	91.9	98.8	98.8	90.7	96.5	97.7	58.1	90.3	93.5	74.6	88.4	91.0
La-SALAD	+1024	**93.0**	98.8	98.8	**94.2**	**98.8**	**100**	**61.3**	90.3	93.5	**74.6**	**88.8**	**91.2**
CricaVPR	10752	91.9	**98.8**	98.8	88.4	94.2	97.7	64.5	90.3	93.5	69.8	**82.2**	85.4
La-CricaVPR	+1024	**93.0**	**98.8**	**98.8**	**94.2**	**98.8**	**98.8**	64.5	**96.8**	**96.8**	**70.2**	82.1	**86.1**

Query **MixVPR** **La-MixVPR**

Fig. 2: Qualitative results from the Amstertime dataset. From left to right: Query, MixVPR Top-1, and La-MixVPR Top-1 (using CAT Fusion). Correct matches are bordered in green; incorrect in red. Additional qualitative results are provided in the Suppl. Material.

62% (6.82 GFLOPs vs. 17.7 GFLOPs), demonstrating that horizontal scaling via multi-modality is a more efficient route to robustness than the vertical scaling of transformers architectures.

Complementary Synergy versus Sequential Retrieval: Finally, we compare our joint approach against sequential re-ranking (Top-100 retrieval followed by cross-modal re-ranking). As shown in Table 8, sequential ranking fails catastrophically (R@1 drops to 1.2%) due to the retrieval bottleneck: if the primary modality misses the ground truth in the initial top-K, the second cannot recover it. Furthermore, candidates that rank highly in one modality do not necessarily possess discriminative features in the other. Joint fusion avoids this by allowing both modalities to mutually disambiguate the search space from the start.

Table 7: Efficiency-Performance Trade-off. We compare current state-of-the-art (SOTA) VPR, namely, CricaVPR, and La-CricaVPR (ours) with smaller backbones and CAT fusion. CricaVPR is a state-of-the-art (SOTA) VPR model with a Vit-L backbone. CricaVPR-Small corresponds to CricaVPR with a ViT-S backbone. La-CricaVPR-Small pairs CricaVPR-Small with BGE-Small (textual encoder). It achieves higher R@1 than the base SOTA model with $\sim 62\%$ fewer total GFLOPs.

Configuration	**Params**	**GFLOPs**			**Dim**	**MSLS-B**		**MSLS-W**		**Amst.-La**	
	(M)	Inf.	Retr.	Total	(d)	R@1	R@5	R@1	R@5	R@1	R@5
CricaVPR	106.8	17.5	0.20	17.7	10k	91.9	**98.8**	88.4	94.2	**64.5**	90.3
CricaVPR-Small	27.2	4.6	0.10	4.8	5k	89.5	95.3	84.9	91.9	54.8	87.1
La-Crica-Small	60.6	6.7	0.12	6.82	6k	**93.0**	97.7	**90.7**	**96.5**	58.1	**93.5**

Table 8: Comparison of Joint Fusion (Ours) vs. Sequential Re-ranking on Amstertime. Sequential methods re-rank the top-100 candidates from the initial search modality.

Initial Search	**Re-ranked by**	**Amstertime (Recall@K)**			
		R@1	R@5	R@10	R@20
Single Modality Baselines					
MixVPR (Visual)	–	35.7	53.1	60.4	65.9
BGE (Textual)	–	9.5	18.0	23.5	32.4
Sequential Re-ranking (Top-100)					
MixVPR	BGE (Text)	1.2	3.4	5.2	6.6
BGE	MixVPR (Vis)	4.4	5.6	5.9	6.3
Joint Multi-modal Fusion					
LaVPR (Ours)	–	**37.0**	**58.2**	**65.6**	**71.6**

5 Cross-Modal Place Recognition ($L \rightarrow V$)

While multi-modal fusion leverages language as a robust auxiliary signal, a more fundamental challenge lies in cross-modal retrieval: identifying a geographic location within a visual database using only a natural language query. This task requires alignment between abstract linguistic descriptions and the concrete, structural visual features that define a specific place.

5.1 Aligning Language and Vision for 'Blind' Localization

Given an image I and a textual description t, let E_v and E_t denote the visual and textual encoders mapping I and t to embeddings $\mathbf{z}_v \in \mathbb{R}^{d_v}$ and $\mathbf{z}_t \in \mathbb{R}^{d_t}$, respectively. The objective is to retrieve the ground-truth image I^* from a database $\mathcal{D}$ that corresponds to a query description t_q. Note that the text description generated from a query image are is solely to retrieve geographically matching, yet visually distinct, reference images from the database, while the query image itself never appears in the search database. Standard vision-language

models are pre-trained to align object-level semantics rather than the fine-grained spatial and structural cues required to discriminate between visually similar locations. To bridge this gap, we do not learn a new latent space from scratch; instead, we re-orient the existing shared embedding space toward place-discriminative features using Low-Rank Adaptation (LoRA) [14]. For a pre-trained weight matrix $W_0 \in \mathbb{R}^{d\times k}$, the weight update is represented as a low-rank decomposition:

$$W = W_0 + \Delta W = W_0 + BA \tag{2}$$

where $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$, with rank $r \ll \min(d, k)$. We apply these updates to all linear projections within the Transformer layers of our vision-language backbones. The model is optimized using the Multi-Similarity (MS) loss (Eq. 1), which encourages a discriminative cross-modal manifold by mining hard positive and negative pairs across modalities. We provide additional training and implementation details in Section 6 of the Suppl. Material.

5.2 Results & Discussion

The Challenge of Zero-Shot Alignment: Results in Table 9 show that standard vision-language foundation models (CLIP [32], BLIP [22], EvaClipv2 [9], SigLIPv2 [36]) exhibit poor zero-shot performance, with R@1 scores frequently below 3%. This underscores a fundamental "semantic-structural gap": generic semantic representations optimized for object-level concepts lack the fine-grained architectural cues essential for localization.

Table 9: Cross-modal Vision-Language Retrieval performance. We compare cross-modal foundation models against our language-augmented (LoRA-MS-) versions. Bold indicates the best result within each model pair.

Model	Amstertime			MSLS-val			Pitts30		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
CLIP	2.0	7.5	12.1	2.3	6.2	10.4	10.9	29.6	43.5
LoRA-MS-CLIP	**11.5**	**27.5**	**37.7**	**38.0**	**58.4**	**68.2**	**49.3**	**74.4**	**82.8**
BLIP	1.3	4.1	7.0	1.6	5.9	8.2	8.1	26.8	41.1
LoRA-MS-BLIP	**12.8**	**31.4**	**41.1**	**38.0**	**60.5**	**68.4**	**50.2**	**74.2**	**82.8**
EVA-CLIP-V2	2.8	8.0	12.7	3.8	9.2	13.0	11.4	30.9	44.2
LoRA-MS-EVA-V2	**11.5**	**27.9**	**36.8**	**32.8**	**52.8**	**62.4**	**41.5**	**66.8**	**76.6**
SigLIP-V2	0.9	3.1	5.5	2.6	6.5	9.7	10.1	30.1	44.3
LoRA-MS-SigLIP-V2	**13.8**	**29.0**	**40.0**	**35.7**	**60.5**	**69.6**	**49.3**	**74.6**	**82.8**

Consistent Efficacy of Parameter-Efficient Fine-Tuning and MS Alignment: We observe a consistent and substantial performance gain across all evaluated architectures when utilizing Parameter-Efficient Fine-Tuning (PEFT) via LoRA. As detailed in Table 9, our LoRA-MS-variants provide an order-of-magnitude improvement over their respective baseline models. For instance, LoRA-MS-SigLIP-V2 reaches an R@1 of 13.8 on Amstertime and 35.7 on MSLS-val, compared to 0.9 and 2.6 for the base model, respectively. Notably, on the Pitts-30k benchmark, LoRA-MS-SigLIP-V2 R@5 performance peaks at 74.6%, rising from a baseline of 30.0%. This trend remains stable across different backbones, confirming that LoRA successfully injects domain-specific VPR knowledge while preventing catastrophic forgetting.

Ablation of Training Strategy: Our training strategy ablation (Table 10, rows 2–5) reveals that the synergy between LoRA and the Multi-Similarity (MS) loss is the primary driver of representation convergence. Standard adaptation techniques prove insufficient: both isolated Learned Pooling and full fine-tuning with the MS loss result in near-total failure (R@1 $\approx$ 0.1%). This collapse occurs because full fine-tuning causes catastrophic forgetting of the foundational VLM's broad semantic priors, which conflict with the rigid visual-discriminative objectives of the VPR task.

By contrast, applying LoRA jointly to both the vision and text towers restricts adaptation to low-rank updates, allowing each encoder to align its representations toward place-discriminative cues while preserving its pre-trained priors. Concurrently, the MS loss provides the hard-pair mining necessary to separate geographically similar locations, a signal that standard contrastive objectives (such as the original BLIP loss) lack, leading to markedly lower performance. Ultimately, coupling the parameter-efficient capacity of LoRA with the hard-mining capabilities of the MS loss is what enables the framework to successfully bridge the semantic-structural gap.

Table 10: Comparison of training strategies for Cross-modal VPR over BLIP model.

Training Strategy	**Loss**	**Amstertime (Recall@K)**			
		R@1	R@5	R@10	R@20
Zero-shot	Frozen Encoders	1.3	4.1	7.0	12.3
LLP	MS	0.1	0.5	1.0	2.4
Full Fine-tuning	MS	0.1	0.4	0.9	1.6
LoRA (All Linear, $r = 64$)	Contrastive	0.1	0.4	0.8	1.6
LoRA (All Linear, $r = 64$)	**MS**	**12.8**	**31.4**	**41.1**	**53.2**

LoRA Configuration and Coverage: The ablation results in Table 11 indicate that the scope of adaptation is as critical as the rank itself. Targeting all linear layers outperforms targeting only Q, K, V matrices. Extending coverage across

all projections with rank $r = 64$ provides the necessary capacity to capture complex environmental semantics, yielding a peak R@1 of 12.8% on Amstertime. This suggests that the structural nuances required for VPR are distributed throughout the Transformer blocks, necessitating a comprehensive adaptation of both attention and feed-forward mechanisms.

Table 11: Ablation of LoRA hyperparameters and Loss functions over BLIP model.

Target Layers	**Rank (r)**	**Loss**	**Amstertime (Recall@K)**			
			R@1	R@5	R@10	R@20
LoRA (QKV)	16	Contrastive	4.0	10.1	13.9	19.9
LoRA (QKV)	16	MS	8.9	25.0	34.0	44.8
LoRA (QKV)	64	MS	10.6	29.1	38.0	49.3
LoRA (All Linear)	64	Contrastive	0.1	0.4	0.8	1.6
LoRA (All Linear)	**64**	**MS**	**12.8**	**31.4**	**41.1**	**53.2**

Impact of Descriptive Density: Table 12, LoRA-MS-BLIP's performance scales directly with narrative length, and consistently outperforms the zero-shot BLIP baseline. Specifically, increasing the average query length from 15 to 60 words yields a 100% relative R@1 gain for LoRA-MS-BLIP (6.5% to 13.2%). In contrast, the zero-shot baseline peaks at 35 words ($R@1 = 3.2\%$) before collapsing to 1.3% as complexity increases. This divergence suggests that while standard foundation models are optimized for short, object-centric captions, our training paradigm enables the model to resolve complex scenes by aggregating fine-grained structural cues distributed throughout a dense, multi-element narrative.

Table 12: Impact of Query Length on Cross-Modal Retrieval. Descriptive statistics are reported in word counts.

Model	**Query Length Statistics**			**Amstertime (Recall@K)**			
	Min	Avg ± Std	Max	R@1	R@5	R@10	R@20
Full Narratives							
BLIP	6	15 ± 1.9	20	2.8	8.8	13.2	19.3
LoRA-MS-BLIP	6	15 ± 1.9	20	6.5	17.1	24.9	34.8
BLIP	10	25 ± 4.9	44	3.1	9.7	13.6	18.6
LoRA-MS-BLIP	10	25 ± 4.9	44	9.0	24.7	34.0	42.1
BLIP	11	35 ± 7.2	74	3.2	7.7	13	19.0
LoRA-MS-BLIP	11	35 ± 7.2	74	11.0	25.9	35.8	45.9
BLIP	10	60 ± 22	184	1.3	4.1	7.0	12.3
LoRA-MS-BLIP	10	60 ± 22	184	13.2	31.8	42.3	52.8

6 Limitations and Future Work

Our framework exhibits a few limitations that highlight promising directions for future research. First, we observe diminishing returns when applying language augmentation to state-of-the-art self-supervised backbones; when underlying visual features are already highly discriminative, the introduced linguistic context can occasionally inject sub-optimal priors. Second, despite the substantial performance gains enabled by our LoRA-MS paradigm, a significant performance gap persists between cross-modal and uni-modal retrieval. While our baseline represents a major leap over zero-shot VLM performance, "blind" localization from text descriptions alone remains markedly less reliable than image-based retrieval. We explicitly highlight closing this cross-modal gap as an open challenge and a call to action for the community.

Finally, to isolate and evaluate the immediate utility of language augmentation, this study prioritized late-fusion architectures. Although our preliminary experiments with mid-fusion configurations resulted in performance degradation, the trade-off between enforcing tight semantic alignment via early-fusion strategies and preserving fine-grained, discriminative visual features remains a complex architectural dynamic. A systematic evaluation of these early vs. late-fusion trade-offs is beyond the scope of this work. Consequently, future research will investigate more sophisticated cross-modal alignment objectives, advanced multi-stage fusion mechanisms, and the integration of additional modalities to bridge these remaining performance gaps.

7 Conclusion

This work introduces LaVPR, a large-scale benchmark that extends vision-only datasets with high-density linguistic descriptions to investigate multi-modal and cross-modal localization. Our analysis demonstrates that language serves as a critical robustness anchor; specifically, our ADS-LLP mechanism yields consistent gains and superior computational efficiency compared to the vertical scaling of visual transformers. Finally, we establish that the synergy between LoRA and Multi-Similarity (MS) loss is essential for effective cross-modal retrieval, providing order-of-magnitude improvements over standard baselines.

References

1. Abdin, M.I., Ade Jacobs, S., Awan, A.A., et al.: Phi-3 technical report: A highly capable language model locally on your phone. Tech. Rep. MSR-TR-2024-12, Microsoft (August 2024), https://arxiv.org/abs/2404.14219
2. Ali-bey, A., Chaib-draa, B., Giguere, P.: Gsv-cities: Toward appropriate supervised visual place recognition. Neurocomputing (2022)
3. Ali-Bey, A., Chaib-Draa, B., Giguere, P.: Mixvpr: Feature mixing for visual place recognition. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 2998–3007 (2023)

4. Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J.: Netvlad: Cnn architecture for weakly supervised place recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5297–5307 (2016)
5. Berton, G., Masone, C., Caputo, B.: Rethinking visual geo-localization for large-scale applications. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4878–4888 (2022)
6. Berton, G., Trivigno, G., Caputo, B., Masone, C.: Eigenplaces: Training viewpoint robust models for visual place recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11080–11090 (2023)
7. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., et al.: Sam 3: Segment anything with concepts. arXiv preprint arXiv:2511.16719 (2025)
8. Carlevaris-Bianco, N., Ushani, A.K., Eustice, R.M.: University of Michigan North Campus long-term vision and lidar dataset. International Journal of Robotics Research **35**(9), 1023–1035 (2015)
9. Fang, Y., Sun, Q., Wang, X., Huang, T., Wang, X., Cao, Y.: Eva-02: A visual representation for neon genesis. Image and Vision Computing p. 105171 (2024)
10. Google: Gemini 2.5 flash (2025), `https://deepmind.google/technologies/gemini/`, large Language Model
11. Hausler, S., Garg, S., Xu, M., Milford, M., Fischer, T.: Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14141–14152 (2021)
12. Hoffer, E., Ailon, N.: Deep metric learning using triplet network. In: International workshop on similarity-based pattern recognition. pp. 84–92. Springer (2015)
13. Hong, Z., Petillot, Y., Lane, D., Miao, Y., Wang, S.: Textplace: Visual place recognition and topological localization through reading scene texts. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2861–2870 (2019)
14. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022)
15. Hu, H., Qiao, Z., Cheng, M., Liu, Z., Wang, H.: Dasgil: Domain adaptation for semantic and geometric-aware image-based localization. IEEE Transactions on Image Processing **30**, 1342–1353 (2020)
16. Huang, G., Zhou, Y., Hu, X., Zhang, C., Zhao, L., Gan, W.: Dino-mix enhancing visual place recognition with foundational vision model and feature mixing. Scientific Reports **14**(1), 22100 (2024)
17. Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., et al.: A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems **43**(2), 1–55 (2025)
18. Izquierdo, S., Civera, J.: Optimal transport aggregation for visual place recognition. In: Proceedings of the ieee/cvf conference on computer vision and pattern recognition. pp. 17658–17668 (2024)
19. Kolmet, M., Zhou, Q., Ošep, A., Leal-Taixé, L.: Text2pos: Text-to-point-cloud cross-modal localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6687–6696 (2022)
20. Komorowski, J., Wysoczańska, M., Trzcinski, T.: Minkloc++: lidar and monocular image fusion for place recognition. In: 2021 International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (2021)

21. Lai, H., Yin, P., Scherer, S.: Adafusion: Visual-lidar fusion with adaptive weights for place recognition. IEEE Robotics and Automation Letters **7**(4), 12038–12045 (2022)
22. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International conference on machine learning. pp. 12888–12900. PMLR (2022)
23. Li, K., Ou, Y., Ning, J., Kong, F., Cai, H., Li, H.: Unified depth-guided feature fusion and reranking for hierarchical place recognition. Sensors **25**(13), 4056 (2025)
24. Li, Z., Shang, T., Xu, P., Deng, Z.: Place recognition meet multiple modalities: a comprehensive review, current challenges and future development. Artificial Intelligence Review **58**(11), 363 (2025)
25. Liu, D., Huang, S., Li, W., Shen, S., Wang, C.: Text to point cloud localization with multi-level negative contrastive learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 5397–5405 (2025)
26. Lu, F., Lan, X., Zhang, L., Jiang, D., Wang, Y., Yuan, C.: Cricavpr: Cross-image correlation-aware representation learning for visual place recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16772–16782 (2024)
27. Lu, F., Zhang, L., Lan, X., Dong, S., Wang, Y., Yuan, C.: Towards seamless adaptation of pre-trained models for visual place recognition. In: The Twelfth International Conference on Learning Representations (2024), `https://openreview.net/forum?id=TVg6hlfsKa`
28. Maddern, W., Pascoe, G., Linegar, C., Newman, P.: 1 year, 1000 km: The oxford robotcar dataset. The International Journal of Robotics Research **36**(1), 3–15 (2017)
29. Melekhin, A., Yudin, D., Petryashin, I., Bezuglyj, V.: Mssplace: Multi-sensor place recognition with visual and text semantics. IEEE Access (2025)
30. Piasco, N., Sidibé, D., Gouet-Brunet, V., Demonceaux, C.: Improving image description with auxiliary modality for visual localization in challenging conditions. International Journal of Computer Vision **129**(1), 185–202 (2021)
31. Radenović, F., Tolias, G., Chum, O.: Fine-tuning cnn image retrieval with no human annotation. IEEE transactions on pattern analysis and machine intelligence **41**(7), 1655–1668 (2018)
32. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
33. Shang, T., Li, Z., Xu, P., Qiao, J., Chen, G., Ruan, Z., Hu, W.: Bridging text and vision: A multi-view text-vision registration approach for cross-modal place recognition. arXiv preprint arXiv:2502.14195 (2025)
34. Torii, A., Arandjelovic, R., Sivic, J., Okutomi, M., Pajdla, T.: 24/7 place recognition by view synthesis. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1808–1817 (2015)
35. Torii, A., Sivic, J., Pajdla, T., Okutomi, M.: Visual place recognition with repetitive structures. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 883–890 (2013)
36. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al.: Siglip 2: Multilingual vision-language encoders with improved semantic understanding. Localization, and Dense Features **6** (2025)

37. Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., Liu, W.: Cosface: Large margin cosine loss for deep face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5265–5274 (2018)
38. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
39. Wang, X., Han, X., Huang, W., Dong, D., Scott, M.R.: Multi-similarity loss with general pair weighting for deep metric learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5022–5030 (2019)
40. Warburg, F., Hauberg, S., Lopez-Antequera, M., Gargallo, P., Kuang, Y., Civera, J.: Mapillary street-level sequences: A dataset for lifelong place recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2626–2635 (2020)
41. Xia, Y., Shi, L., Ding, Z., Henriques, J.F., Cremers, D.: Text2loc: 3d point cloud localization from natural language. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14958–14967 (2024)
42. Xiao, S., Liu, Z., Zhang, P., Wang, N., , et al.: C-pack: Packged resources for general chinese embedding. arXiv preprint arXiv:2309.07597 (2023)
43. Xie, J., Kiefel, M., Sun, M.T., Geiger, A.: Semantic instance annotation of street scenes by 3d to 2d label transfer. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 3688–3697 (2016)
44. Xu, Y., Qu, H., Liu, J., Zhang, W., Yang, X.: Cmmloc: Advancing text-to-pointcloud localization with cauchy-mixture-model based framework. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 6637–6647 (2025)
45. Yildiz, B., Khademi, S., Siebes, R.M., Van Gemert, J.: Amstertime: A visual place recognition benchmark dataset for severe domain shift. In: 2022 26th International Conference on Pattern Recognition (ICPR). pp. 2749–2755. IEEE (2022)
46. Zhou, Z., Xu, J., Xiong, G., Ma, J.: Lcpr: A multi-scale attention-based lidar-camera fusion network for place recognition. IEEE Robotics and Automation Letters **9**(2), 1342–1349 (2023)
47. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In: The Twelfth International Conference on Learning Representations (2024), `https://openreview.net/forum?id=1tZbq88f27`