TEXT-ASSISTED EVENT STREAM VISUAL PLACE RECOGNITION VIA MULTI-SCALE SPATIAL-TEMPORAL OPTIMAL TRANSPORT

Anonymous authors
Paper under double-blind review

ABSTRACT

Benefiting from the high dynamic range, high temporal resolution, and low energy consumption of event cameras, low-altitude economic research leveraging event cameras has become a popular topic. Several existing attempts introduce natural-language descriptions into visual place recognition (VPR) to mitigate the sparsity of structural spatial information inherent to event cameras. Nevertheless, these methods still fail to fully address two core challenges: effectively encoding event streams and semantically aligning event streams with text. In this paper, we propose a multi-scale spatial-temporal optimal transport approach, MSTOT-EPR, to address these challenges. Specifically, given event streams and scene text information, we first adopt a visual encoder and a text encoder to extract respective feature representations. We then apply optimal-transport constraints to achieve cross-modal alignment. We then design a cross-modal bottleneck fusion module to further fuse bimodal features. Furthermore, we develop multi-scale spatial optimal transport and inter-frame temporal optimal transport modules to mine high-quality spatial-temporal event features for scene recognition. Extensive experiments on the NYC-Event-VPR and EPRBench datasets fully validate the effectiveness of our proposed framework for event-based VPR. The source code of this paper will be released upon acceptance.

1 INTRODUCTION

Visual Place Recognition (VPR) Chen et al. (2026b) aims to match a query image to its corresponding location in a large-scale geotagged database. It holds great research value for many real-world applications, including autonomous driving navigation, large-scale map reconstruction, and low-altitude economy. State-of-the-art VPR solutions are mostly implemented with visible-light cameras; thus, they inherit various inherent challenges from conventional visual perception, such as extreme viewpoint variation, illumination fluctuation, adverse weather conditions, long-term temporal shift, and frequent dynamic occlusions. Developing high-performance VPR approaches remains a challenging problem worthy of further investigation.

Benefiting from the unique advantages of event cameras, including low energy consumption, high dynamic range, and high temporal resolution, event-based vision has attracted increasing attention from researchers Gallego et al. (2020a); Chen et al. (2026a); Wang et al. (2026a). Each pixel of an event camera works independently and records the intensity changes of scenes continuously. Specifically, an individual event is formulated as (x, y, t, p), where x, y denote the spatial pixel coordinates, t represents the timestamp, and p refers to the event polarity. The polarity value of +1 indicates an increase in brightness, while -1 corresponds to a decrease in brightness. In recent years, several studies have begun to explore event-based VPR Wang et al. (2026b); Kong et al. (2022), such as Sparse-Event-VPR Fischer & Milford (2022), which performs training-free place matching using sparse event-count representations, and EventGeM Hines et al. (2026), which combines regional GeM-based global retrieval with local feature matching and RANSAC-based homography re-ranking.

Despite this progress, these works are still limited by the following issues: 1). Existing event stream-based visual place recognition algorithms mostly adopt event representations from a single time

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

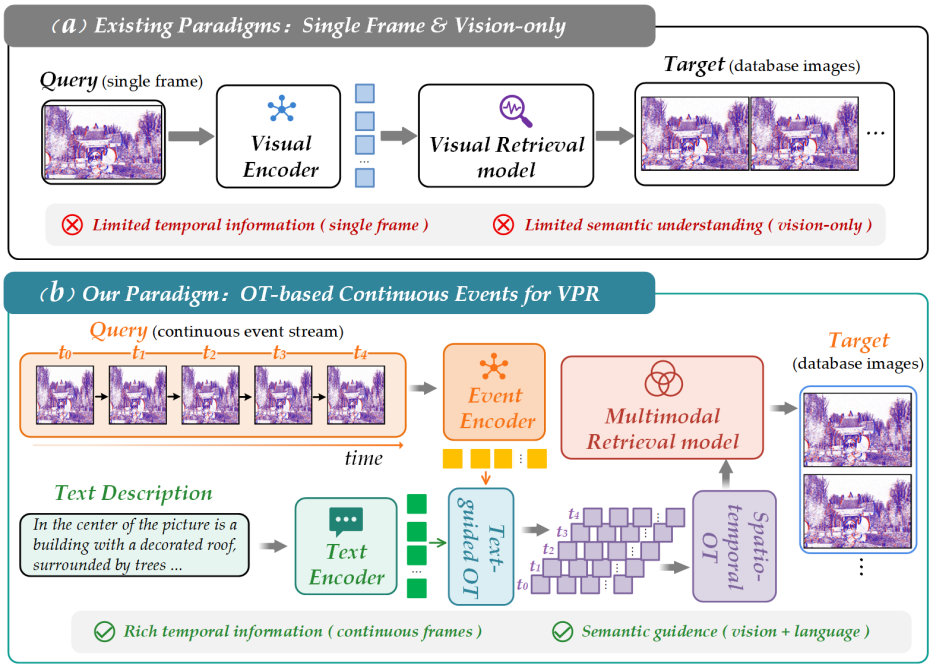


Figure 1: Comparison between conventional vision-only VPR and our event-text VPR paradigm.

window and rarely consider the asynchronous nature of event cameras. Consequently, they are vulnerable to motion-induced positional shifts and instabilities stemming from event sparsity. 2). Few existing works take into account the lack of semantic information caused by the spatial sparsity of event streams, while the visual place recognition task heavily relies on both spatial-detail features and semantic features. 3). Existing text-enhanced methods for event streams adopt coarse-grained modeling strategies, which fail to establish fine alignment between the temporal features of events and the deep semantic information of texts. These methods cannot effectively adapt to the semantic modeling of complex event streams, thereby severely limiting the performance upper bound of event stream understanding and reasoning tasks.

To handle these aforementioned issues, in this paper, we propose a novel multi-scale spatial-temporal optimal transport (OT) based framework for event stream-based VPR, termed MSTOT-EPR. Optimal transport (OT) was originally proposed to find the minimum-cost mapping between two probability distributions, resolving the problem of measuring discrepancy for heterogeneous mass distributions. For spatial-temporal modeling, OT excels at capturing structural correspondences across sequential states rather than simple pointwise differences, making it robust against misalignment and spatial-temporal shift. Specifically, given the event streams and scene text information, we first adopt a vision encoder and a text encoder to extract their feature representations. Then, we apply optimal transport constraints to achieve cross-modal alignment. More importantly, we design the cross-modal bottleneck fusion module to achieve bimodal feature fusion and develop multi-scale spatial optimal transport and inter-frame temporal optimal transport modules to mine high-quality spatial-temporal event features for scene recognition. An overview of our framework can be found in Fig. 2.

To sum up, the main contributions of this paper can be summarized as follows:

1). We propose a novel multi-scale spatial-temporal optimal transport-based framework for event stream-based visual place recognition, termed MSTOT-EPR. The spatial multi-scale OT and temporal inter-frame OT modules are designed to construct feature-and-position-aware cross-scale transport plans and anchor-guided token-level transport plans with gated temporal aggregation, respectively.

2). We introduce the scenario text as an auxiliary modality and further propose a new optimal transport-based cross-modal alignment module to augment the semantic information of event streams.

3). Extensive experiments on multiple event-based VPR benchmark datasets fully validated the effectiveness of our proposed framework for the representation learning of text-assisted event streams.

2 Related Work

Visual Place Recognition. Visual Place Recognition (VPR) is a fundamental capability for autonomous navigation and spatial AI, aiming to identify the geographical location of a query observation by retrieving matching candidates from a reference database. Early learning-based approaches mainly focused on learning compact and discriminative global descriptors from RGB images. NetVLAD Arandjelovic et al. (2016) introduced a differentiable VLAD layer for weakly supervised place recognition, while subsequent methods such as ConvAP Ali-bey et al. (2022) and MixVPR Ali-Bey et al. (2023) further improved feature aggregation and descriptor robustness. With the development of transformer architectures, ViT-based models such as TransVPR Wang et al. (2022) have enhanced long-range dependency modeling and feature alignment for challenging scenes. Recently, foundation models and query-based aggregation strategies have further advanced VPR. AnyLoc Keetha et al. (2023) demonstrates that off-the-shelf self-supervised visual features can provide strong zero-shot place recognition performance. BoQ Ali-bey et al. (2024) introduces a bag-of-queries mechanism to aggregate patch-level features into compact global descriptors softly. SALAD Mahjoub & Kasaei (2023) explores optimal-transport-inspired aggregation for dynamic feature clustering, improving robustness under challenging appearance variations.

Despite these advances, most existing VPR methods are designed for synchronous RGB frames and rely heavily on dense appearance cues. Their performance can degrade under low illumination, high dynamic range, high-speed motion, and severe motion blur. Moreover, existing aggregation strategies mainly operate on spatial image patches, leaving the sparse and asynchronous temporal structure of event streams insufficiently explored. In contrast, our work focuses on event-stream-based VPR and introduces a spatial-temporal alignment framework tailored to the unique characteristics of event data.

Event-based and Multimodal Perception. Event cameras are bio-inspired asynchronous sensors that respond to per-pixel log-intensity changes, providing microsecond-level temporal resolution, high dynamic range, low latency, and resilience to motion blur Gallego et al. (2020b). These properties make event streams attractive for robust visual perception in extreme scenarios where conventional RGB cameras may fail. Early event-based methods usually converted raw events into frame-like representations, such as event histograms, time surfaces, or voxel grids, enabling the use of conventional convolutional architectures Rebecq et al. (2019). Later studies explored graph-based, point-based, and spiking neural models to better preserve the sparse and asynchronous nature of event data Bi et al. (2019). In the context of place recognition, EventVLAD Fischer & Milford (2020) and recent event-based VPR benchmarks Lee et al. (2023) have shown the potential of event cameras for robust localization under challenging visual conditions. However, event streams mainly encode brightness changes rather than absolute appearance, making them semantically sparse and difficult to interpret compared with RGB images. To compensate for limited semantic information, multimodal learning has become an effective strategy. Vision-language pre-training models such as CLIP Radford et al. (2021) map visual and textual information into a shared representation space, enabling strong cross-modal generalization. Text-guided and language-assisted recognition methods further demonstrate that natural language descriptions can provide high-level semantic priors for visual understanding Zhou et al. (2022); Liang et al. (2022). In VPR, language descriptions can help disambiguate visually similar places by introducing scene-level contextual cues.

Nevertheless, directly applying vision-language fusion to event-based VPR remains challenging. Event features are sparse, asynchronous, and temporally ordered, whereas textual features are discrete and semantic. Simple concatenation, projection, or contrastive matching may ignore the fine-grained spatial correspondence and temporal evolution of event streams. Therefore, an effective text-assisted event VPR framework should explicitly align multimodal representations across both spatial and temporal dimensions. This motivates our multi-scale spatial-temporal optimal transport design.

Optimal Transport in Computer Vision. Optimal transport (OT) establishes soft correspondences between probability distributions by minimizing the cost of transporting mass under marginal constraints Villani (2009); Peyré & Cuturi (2019). Unlike pointwise similarity or global pooling, OT

captures distribution-level structure and supports many-to-many matching, making it suitable for aligning heterogeneous visual representations. Entropic regularization further enables efficient and differentiable optimization Cuturi (2013), facilitating applications in domain adaptation and structured feature matching Courty (2017); Sarlin (2020). Existing OT-based VPR methods, such as SALAD Mahjoub & Kasaei (2023), primarily focus on spatial aggregation from static RGB images. In contrast, our approach extends OT to event-text alignment, multi-scale spatial correspondence, and inter-frame temporal modeling. Further background and a detailed discussion of OT-related vision methods are provided in Appendix A.

3 METHODOLOGY

3.1 OVERVIEW

Event streams are robust to illumination changes and rapid motion, but their spatial sparsity limits the extraction of discriminative place representations. Textual descriptions provide complementary scene semantics, yet the modality gap between event and text features, together with the temporal variations across event frames, remains challenging to model. To address these issues, we propose Multi-Scale spatial-temporal Optimal Transport for Event-stream Place Recognition (MSTOT-EPR), which establishes structured correspondences across modalities, spatial scales, and time using optimal transport. As illustrated in Fig. 2, we first extract event and text features and align them through cross-modal OT and bottleneck fusion. Multi-scale spatial OT then captures spatial structures at different resolutions, while inter-frame temporal OT aligns consecutive frames to a shared anchor support. Finally, the enhanced spatial-temporal features are aggregated into a compact global descriptor for place retrieval.

3.2 INPUT REPRESENTATIONS

Visual representation. Given an event stream $\mathcal{E} = \{(x_i, y_i, t_i, p_i)\}_{i=1}^{N}$, where (x_i, y_i), t_i, and $p_i \in \{-1, +1\}$ denote the spatial location, timestamp, and polarity of the i-th event, respectively, we partition it into T temporal segments and construct an event-frame sequence $\{\mathbf{I}_t\}_{t=0}^{T}$. For each frame $\mathbf{I}_t \in \mathbb{R}^{H\times W}$, the DINOv2 visual encoder Oquab et al. (2024) extracts a global CLS token $\mathbf{c}_t \in \mathbb{R}^{C_v}$ and a set of patch tokens $\mathbf{V}_t = [\mathbf{v}_{t,1}, \ldots, \mathbf{v}_{t,M}] \in \mathbb{R}^{M\times C_v}$, where $M = HW/p^2$ and p is the patch size. The CLS token summarizes frame-level content, while the patch tokens preserve local spatial structure.

Textual representation. For the paired scene description $\mathcal{T}$, the CLIP text encoder Radford et al. (2021) produces a sentence embedding $\mathbf{s} \in \mathbb{R}^{C_s}$, word-token features $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_L] \in \mathbb{R}^{L\times C_w}$, and an attention mask $\mathbf{m} \in \{0, 1\}^L$. We discard special and padded tokens according to $\mathbf{m}$, retaining L' valid word tokens. The sentence embedding provides global semantic supervision, whereas the valid word tokens support fine-grained cross-modal alignment.

3.3 OT-ALIGNED MODALITY FUSION

To bridge the representation gap between event and text modalities, we first establish fine-grained cross-modal correspondences via optimal transport and then integrate the aligned features through bottleneck-based fusion. Specifically, we first project the visual patch tokens $\mathbf{V}_t$ and word tokens $\mathbf{W}$ into a shared D-dimensional space using two learnable projections g_v and g_w. Their ℓ_2-normalized representations are denoted by $\hat{\mathbf{v}}_{t,i} = \mathrm{Norm}(g_v(\mathbf{v}_{t,i}))$ and $\hat{\mathbf{w}}_j = \mathrm{Norm}(g_w(\mathbf{w}_j))$, respectively. To quantify the local discrepancy between the i-th visual patch of frame t and the j-th valid word token, we define their pairwise matching cost using cosine distance:

$$\mathbf{C}_{t,ij}^{loc} = 1 - \hat{\mathbf{v}}_{t,i}^{\top}\hat{\mathbf{w}}_j. \tag{1}$$

where the pairwise costs are collected into the cost matrix $\mathbf{C}_t^{\mathrm{loc}}$. Based on the cost matrix, we formulate the fine-grained cross-modal alignment as the following entropically regularized optimal transport problem:

$$\mathbf{P}_t^{loc} = \arg\min_{\mathbf{P}\in\mathcal{U}(\boldsymbol{\mu},\boldsymbol{\nu})} \langle \mathbf{P}, \mathbf{C}_t^{loc}\rangle + \varepsilon \sum_{i,j} P_{ij}(\log P_{ij} - 1), \tag{2}$$

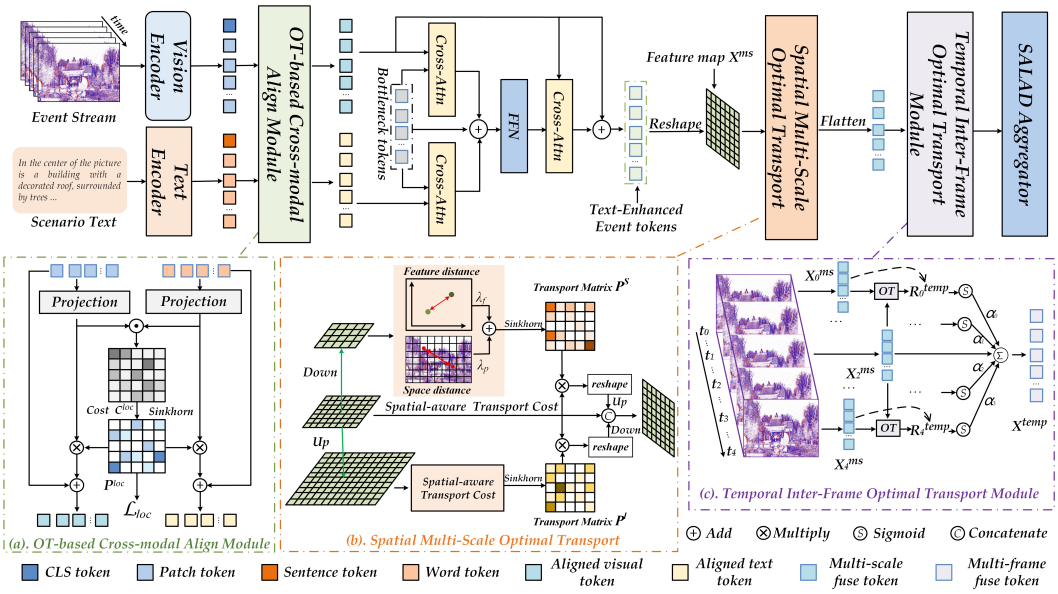


Figure 2: An overview of our proposed text-assisted event stream visual place recognition via multi-scale spatial-temporal optimal transport, termed MSTOT-EPR. The framework integrates OT-based cross-modal alignment with bottleneck fusion, multi-scale spatial-aware OT, and temporal inter-frame OT, followed by SALAD-based feature aggregation.

where $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ assign uniform mass to the visual patches and valid word tokens, respectively, i.e., $\mu_i = 1/M$ and $\nu_j = 1/L'$. $\varepsilon > 0$ controls the strength of the entropic regularization.

We solve equation 2 using the numerically stable log-domain Sinkhorn algorithm (Cuturi, 2013; Schmitzer, 2019). The detailed iterative updates are provided in Appendix B.

For brevity, we denote the resulting transport plan as

$$\mathbf{P}_t^{\mathrm{loc}} = \mathrm{Sinkhorn}\left(\mathbf{C}_t^{\mathrm{loc}}; \boldsymbol{\mu}, \boldsymbol{\nu}, \epsilon\right). \tag{3}$$

Then, the resulting transport matrix $\mathbf{P}_t^{loc}$ is used to obtain bidirectionally aligned representations through barycentric projection, such as:

$$\mathbf{V}_t^{ot} = \hat{\mathbf{V}}_t + \mathrm{Diag}(\boldsymbol{\mu})^{-1}\mathbf{P}_t^{loc}\hat{\mathbf{W}}, \tag{4}$$

$$\mathbf{W}_t^{ot} = \hat{\mathbf{W}} + \mathrm{Diag}(\boldsymbol{\nu})^{-1}(\mathbf{P}_t^{loc})^{\top}\hat{\mathbf{V}}_t. \tag{5}$$

where $\mathrm{Diag}(\cdot)$ denotes the operation that constructs a diagonal matrix from its input vector. Here, $\hat{\mathbf{V}}_t$ and $\hat{\mathbf{W}}$ collect normalized visual and word embeddings, respectively. Thus, $\mathbf{V}_t^{ot} \in \mathbb{R}^{M\times D}$ enriches each visual patch with its textual context, while $\mathbf{W}_t^{ot} \in \mathbb{R}^{L'\times D}$ injects the corresponding visual context into each valid word token.

The OT-aligned representations are subsequently fused using a small set of learnable bottleneck tokens. Let $\mathbf{B} \in \mathbb{R}^{N_b\times D}$ denote the bottleneck tokens, where $N_b \ll M, L'$. They first attend to the aligned visual and textual tokens independently as:

$$\mathbf{B}_v = \mathrm{Attn}(\mathbf{B}, \mathbf{V}_t^{ot}, \mathbf{V}_t^{ot}), \quad \mathbf{B}_w = \mathrm{Attn}(\mathbf{B}, \mathbf{W}_t^{ot}, \mathbf{W}_t^{ot}). \tag{6}$$

where $\mathrm{Attn}(\cdot)$ denotes the multi-head cross-attention operation. Then, the modality-specific representations are integrated with the original bottleneck tokens through residual aggregation, followed by a feed-forward network:

$$\tilde{\mathbf{B}} = \mathrm{FFN}(\mathbf{B} + \mathbf{B}_v + \mathbf{B}_w). \tag{7}$$

The updated bottleneck tokens then broadcast the fused information back to the visual tokens:

$$\tilde{\mathbf{V}}_t = \mathbf{V}_t + g_o(\mathbf{V}_t^{ot} + \mathrm{Attn}(\mathbf{V}_t^{ot}, \tilde{\mathbf{B}}, \tilde{\mathbf{B}})). \tag{8}$$

3.4 SPATIAL MULTI-SCALE OPTIMAL TRANSPORT

To capture spatial structures at multiple resolutions, we construct multi-scale feature representations and align them via spatial-aware optimal transport. Specifically, we first reshape the fused visual tokens $\tilde{\mathbf{V}}_t$ into the event feature map $\tilde{\mathbf{X}}_t \in \mathbb{R}^{C_v \times H/p \times W/p}$. Then, we derive three feature branches from $\tilde{\mathbf{X}}_t$, corresponding to small, middle, and large spatial scales as follows:

$$\mathbf{F}_s = \mathrm{Down}(\tilde{\mathbf{X}}_t), \quad \mathbf{F}_m = \phi(\tilde{\mathbf{X}}_t), \quad \mathbf{F}_l = \mathrm{Up}(\tilde{\mathbf{X}}_t), \tag{9}$$

where $\mathbf{F}_s$ and $\mathbf{F}_l$ are resized to 8×8 and 32×32, respectively, and $\phi(\cdot)$ is a 1×1 projection for the middle branch. We apply two spatial-aware OT operations to align the small- and large-scale branches with the middle branch. Since both operations share the same procedure, we detail only the small-to-middle alignment. Let $\{\mathbf{f}_i^s\}_{i=1}^{M_s}$ and $\{\mathbf{f}_j^m\}_{j=1}^{M}$ denote the flattened small- and middle-scale tokens. Their corresponding normalized spatial coordinates are denoted by $\{\mathbf{r}_i^s\}_{i=1}^{M_s}$ and $\{\mathbf{r}_j^m\}_{j=1}^{M_m}$. We define the spatially aware transport cost by jointly considering the feature discrepancy and spatial distance as:

$$\mathbf{C}_{t,ij}^s = \lambda_f \|\mathbf{f}_i^s - \mathbf{f}_j^m\|_2 + \lambda_p \|\mathbf{r}_i^s - \mathbf{r}_j^m\|_2, \tag{10}$$

where we fix $\lambda_f = 0.9$ and $\lambda_p = 0.1$ in all experiments. Based on this cost matrix, we formulate the small-to-middle cross-scale alignment as an entropically regularized optimal transport problem and solve it by the same log-domain Sinkhorn iterations as introduced in Appendix B as:

$$\mathbf{P}_t^s = \mathrm{Sinkhorn}(\mathbf{C}_t^s; \boldsymbol{\mu}_s, \boldsymbol{\nu}_m, \varepsilon_s), \tag{11}$$

where $\boldsymbol{\mu}_s$ and $\boldsymbol{\nu}_m$ are the uniform source and target marginals, assigning equal transport mass to all tokens at the small and middle scales, respectively. The parameter ϵ_s is the entropic regularization coefficient, which is set to 0.08 to obtain a smooth and numerically stable transport plan. The middle-scale tokens are transported onto the small-scale support and then interpolated back to the middle resolution:

$$\hat{\mathbf{F}}_s = \mathrm{Up}\big(\mathrm{Reshape}\big(\mathbf{P}_t^s \mathbf{F}_m^{tok}\big)\big), \tag{12}$$

where $\mathbf{F}_m^{tok} \in \mathbb{R}^{M \times C_v}$ denotes the flattened middle-scale tokens.

Similarly, for the large-to-middle branch, we first construct the cost matrix $\mathbf{C}_t^l$ and then compute the corresponding transport plan $\mathbf{P}_t^l$ following the same procedure. The middle-scale tokens are subsequently transported onto the large-scale support, and the resulting features are downsampled to obtain $\hat{\mathbf{F}}_l$. Finally, the aligned small- and large-scale features are concatenated with the original middle-scale feature to form the multi-scale representation:

$$\mathbf{X}_t^{ms} = \mathrm{Concat}(\hat{\mathbf{F}}_s, \mathbf{F}_m, \hat{\mathbf{F}}_l). \tag{13}$$

The resulting feature map $\mathbf{X}_t^{\mathrm{ms}} \in \mathbb{R}^{3C_v \times H/p \times W/p}$ integrates spatial information across multiple resolutions and is subsequently fed into the temporal inter-frame OT module.

3.5 TEMPORAL INTER-FRAME OPTIMAL TRANSPORT

Event streams contain rich temporal dynamics that cannot be fully represented by a single frame. We therefore introduce temporal inter-frame OT to align the spatially enhanced feature maps to a fixed temporal anchor. Let $\mathbf{X}_t^{ms}$ denote the spatial-OT output for frame t, and let $\mathbf{F}_t = [\mathbf{f}_{t,1}, \ldots, \mathbf{f}_{t,M'}]^\top \in \mathbb{R}^{M' \times C'}$ be its flattened token matrix, where the lowercase vector $\mathbf{f}_{t,i} \in \mathbb{R}^{C'}$ denotes the i-th spatial token of frame t. For each five-frame input sequence, we select the middle frame (indexed by $a = 2$) as the temporal anchor during both training and inference. Accordingly, $\mathbf{F}_a$ denotes the anchor-frame token matrix and $\mathbf{f}_{a,i}$ its i-th token. For each frame $t \neq a$, we compute the pairwise temporal matching cost between the anchor tokens and the current-frame tokens. Both token features are first ℓ_2-normalized, and the cost is defined as the cosine distance:

$$\mathbf{C}_{t,ij}^{temp} = 1 - \frac{\mathbf{f}_{a,i}^\top \mathbf{f}_{t,j}}{\|\mathbf{f}_{a,i}\|_2 \|\mathbf{f}_{t,j}\|_2}. \tag{14}$$

Here, $\mathbf{C}_t^{temp} \in \mathbb{R}^{M' \times M'}$ compares every anchor token $\mathbf{f}_{a,i}$ (row i) with every token $\mathbf{f}_{t,j}$ from frame t (column j). Using uniform marginals over the two token sets, we compute the temporal transport plan as

$$\mathbf{P}_t^{\mathrm{temp}} = \mathrm{Sinkhorn}\left(\mathbf{C}_t^{\mathrm{temp}}; \boldsymbol{\mu}_a, \boldsymbol{\nu}_t, \varepsilon_{\mathrm{temp}}\right), \ t \neq a, \tag{15}$$

which follows the same log-domain Sinkhorn procedure introduced in Appendix B. Here, $\boldsymbol{\mu}_a$ and $\boldsymbol{\nu}_t$ are the uniform marginals that assign equal transport mass to all tokens in the anchor frame and the current frame, respectively. The parameter ϵ_{temp} is the entropic regularization coefficient, which is set to 0.08. We perform 20 Sinkhorn iterations to obtain the temporal transport plan.

The transport plan is row-normalized and subsequently used to transport the current-frame tokens onto the anchor-frame support:

$$\widetilde{\mathbf{F}}_t = \text{Diag}\left(\mathbf{P}_t^{\text{temp}}\mathbf{1}\right)^{-1}\mathbf{P}_t^{\text{temp}}\mathbf{F}_t, t \neq a. \tag{16}$$

For the anchor itself, we use the identity correspondence, i.e., $\widetilde{\mathbf{F}}_a = \mathbf{F}_a$. This produces a temporally aligned token sequence $\{\widetilde{\mathbf{F}}_t\}_{t=1}^T$, all represented on the anchor-frame token support.

To assign different importance to different temporal observations, we first average the tokens of each aligned frame and predict a scalar temporal logit with a lightweight MLP:

$$\boldsymbol{h}_t = \frac{1}{M'}\sum_{i=1}^{M'}\widetilde{\mathbf{f}}_{t,i},\ \alpha_t = \text{Softmax}_{t=1}^{T}\big(\text{MLP}(\boldsymbol{h}_t)\big). \tag{17}$$

The weighted temporal context is then combined with the anchor representation using a learnable sigmoid gate γ:

$$\mathbf{F}^{temp} = \mathbf{F}_a + \gamma\left(\sum_{t=1}^{T}\alpha_t\widetilde{\mathbf{F}}_t - \mathbf{F}_a\right),\ \gamma = \sigma(\gamma_0). \tag{18}$$

where γ_0 is a learnable scalar initialized to -2.0, and $\gamma = \sigma(\gamma_0)$ adaptively balances the anchor feature and the aggregated temporal context. The same temporal weights are applied to the CLS tokens when available, yielding a temporally fused global token $\mathbf{c}^{temp}$. The fused token matrix $\mathbf{F}^{temp}$ is reshaped into a feature map $\mathbf{X}^{temp} \in \mathbb{R}^{C'\times H'\times W'}$. The pair $(\mathbf{X}^{temp}, \mathbf{c}^{temp})$ is then passed to the standalone SALAD aggregation layer Mahjoub & Kasaei (2023). SALAD performs its learnable cluster assignment and residual aggregation on the temporally fused visual features and returns the final compact place descriptor:

$$\mathbf{d} = \text{SALAD}(\mathbf{X}^{temp}, \mathbf{c}^{temp}). \tag{19}$$

The descriptor $\mathbf{d}$ is ℓ_2-normalized by SALAD and is used directly for place retrieval and the retrieval loss during training.

3.6 Training Objective

We optimize MSTOT-EPR using three complementary objectives. The retrieval loss $\mathcal{L}_{\text{retr}}$ learns discriminative place descriptors using the Multi-Similarity loss Wang et al. (2019). The local alignment loss $\mathcal{L}_{\text{loc}}$ minimizes the transport cost between event patches and textual tokens, where gradients are not propagated through the transport plan. The global alignment loss $\mathcal{L}_{\text{mmd}}$ reduces the distribution discrepancy between the event CLS features and sentence embeddings using multi-kernel MMD.

The complete training objective is

$$\mathcal{L} = \mathcal{L}_{\text{retr}} + \lambda\mathcal{L}_{\text{loc}} + \mu\mathcal{L}_{\text{mmd}}. \tag{20}$$

Detailed definitions of the three losses are provided in Appendix C.

4 Experiments

4.1 Dataset and Evaluation Metric

Datasets. We evaluate the proposed method on two event-based VPR benchmarks: **NYC-Event-VPR** Pan et al. (2025) and **EPRBench** Wang et al. (2026b). NYC-Event-VPR provides synchronized RGB images and event streams; we therefore report results on both modalities to assess performance across conventional and event-based observations. EPRBench provides scene-level semantic annotations for event data and includes challenging environmental conditions and substantial appearance variations, making it well suited for evaluating text-assisted event-based place recognition.

Table 1: Comparison with state-of-the-art visual place recognition methods on NYC-Event-VPR-Event, NYC-Event-VPR-RGB, and EPRBench. Results are reported as Recall@K , where $K \in \{1, 5, 10\}$. The best result in each column is highlighted in bold.

Method	Source	NYC-Event-VPR-Event			NYC-Event-VPR-RGB			EPRBench		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
01. Deep visual geo-loc Berton et al. (2022b)	CVPR'22	25.6	41.9	49.1	84.9	89.3	90.4	63.2	77.9	82.8
02. CosPlace Berton et al. (2022a)	CVPR'22	32.1	48.1	55.3	85.4	88.8	89.7	73.2	88.9	92.0
03. MixVPR Ali-Bey et al. (2023)	WACV'23	45.7	63.0	68.9	88.3	91.0	91.5	90.2	93.7	95.6
04. EigenPlaces Berton et al. (2023)	ICCV'23	28.5	43.4	51.2	81.4	91.7	**93.9**	81.1	86.0	89.0
05. R2Former Xu et al. (2023)	CVPR'23	61.2	75.0	78.5	86.3	88.9	89.9	80.0	94.8	96.7
06. CricaVPR Lu et al. (2024a)	CVPR'24	30.3	48.2	55.0	88.3	92.2	92.9	92.6	95.3	96.5
07. SelaVPR Lu et al. (2024b)	ICLR'24	43.3	57.6	62.9	87.8	91.3	91.7	89.9	93.6	95.2
08. EMVP Qiu et al. (2024)	NeurIPS'24	45.7	64.1	70.9	85.5	90.4	91.4	88.7	93.9	95.7
09. SALAD Mahjoub & Kasaei (2023)	CVPR'24	39.6	57.8	64.1	89.1	91.8	92.3	91.3	93.4	95.3
10. SuperVLAD Lu et al. (2024c)	NeurIPS'24	39.7	59.2	66.6	88.8	91.5	92.4	93.3	95.0	96.2
11. BoQ Ali-bey et al. (2024)	CVPR'24	38.9	56.7	63.6	83.2	88.1	89.1	93.1	94.7	95.7
12. ImAge Lu et al. (2025b)	NeurIPS'25	42.4	59.0	65.0	85.7	90.4	91.0	76.2	89.7	92.7
13. PRGS Zuo et al. (2025)	PR'25	43.8	62.1	69.3	82.8	88.6	89.9	87.7	93.9	96.9
14. Pair-VPR Hausler & Moghadam (2025)	RA-L'25	64.0	75.2	78.4	89.4	91.9	92.4	91.9	95.7	95.8
15. FOL Wang et al. (2025)	AAAI'25	55.5	68.2	72.5	85.1	89.3	89.9	93.8	96.0	97.1
16. SelaVPR++ Lu et al. (2025a)	TPAMI'25	37.4	56.3	62.4	66.4	76.6	78.5	90.3	94.3	96.4
17. UniPR-3D Deng et al. (2026)	ECCV'26	44.5	59.0	63.9	69.8	78.2	80.0	91.7	94.9	96.3
18. SAGE Chen et al. (2026b)	ICLR'26	54.9	71.2	76.7	**89.7**	**92.5**	93.0	83.9	90.0	92.5
19. A2GC Li & Shang (2025)	CVPR'26	54.9	70.7	73.4	71.1	78.5	80.3	92.8	95.3	96.8
20. SG-VPR Wang et al. (2026b)	TCSVT'26	57.5	74.1	78.9	86.5	90.9	92.0	94.3	96.1	97.1
21. MSTOT-EPR (Ours)	-	**65.0**	**78.9**	**80.9**	81.5	89.7	90.7	**95.0**	**98.5**	**98.8**

Evaluation metrics. Following the official evaluation protocols of both datasets, we adopt Recall@K (R@K) as the evaluation metric, with $K \in \{1, 5, 10\}$. For each query, database samples are ranked according to the similarity between their predicted descriptors. A query is considered correctly localized if at least one ground-truth reference appears among the top-K retrieved candidates. R@K is the percentage of correctly localized queries over the entire query set; higher values indicate better retrieval performance.

4.2 Implementation Details

We use pretrained DINOv2 ViT-B/14 as the visual backbone and frozen CLIP ViT-L/14 for semantic supervision. Each input consists of five consecutive event frames resized to 224×224, with the middle frame ($a = 2$) selected as the temporal anchor. Unless otherwise specified, the bottleneck dimension is 256, and all OT modules perform 20 Sinkhorn iterations with an entropic regularization coefficient of 0.08. SALAD uses 64 clusters to produce an 8448-dimensional descriptor. The model is trained for 20 epochs using AdamW with an initial learning rate of 6×10^{-5}. Complete implementation details are provided in Appendix D.

4.3 Comparison with State-of-the-Art Methods

Table 1 compares MSTOT-EPR with representative state-of-the-art VPR methods under three evaluation settings. MSTOT-EPR achieves the best performance on both event-based benchmarks, obtaining 65.0%, 78.9%, and 80.9% R@1/R@5/R@10 on NYC-Event-VPR-Event and 95.0%, 98.5%, and 98.8% on EPRBench. These results demonstrate the benefit of jointly modeling event-text semantics, multi-scale spatial structure, and inter-frame temporal correspondence. On the RGB split, MSTOT-EPR remains competitive, although it is not specifically optimized for dense RGB appearance cues. A detailed dataset-wise comparison is provided in Appendix E.

4.4 Ablation Study

Effectiveness of key modules. Table 2 evaluates the three core components on a SALAD-based baseline. Spatial-OT improves R@1 from 91.3% to 92.3%, while adding Temporal-OT or bottleneck fusion further raises it to 93.9% and 93.5%, respectively. Combining all three components achieves the best results of 95.0% R@1, 98.5% R@5, and 98.8% R@10, demonstrating their complementary contributions to spatial, temporal, and semantic modeling.

Hyperparameter analysis. We further investigate the bottleneck dimension, alignment-loss weights, and temporal anchor-frame selection. The complete results are reported in Appendix F.

Table 2: Ablation study of the proposed components on the EPRBench dataset. Spatial optimal transport (Spa-OT), temporal optimal transport (Temp-OT), and bottleneck fusion (BF) are progressively introduced to evaluate their individual and complementary contributions.

Key Modules			Evaluation Metrics		
Spa-OT	**Temp-OT**	**BF**	**R1 (%) ↑**	**R5 (%) ↑**	**R10 (%) ↑**
×	×	×	91.3	93.4	95.3
✓	×	×	92.3	94.6	96.6
✓	✓	×	93.9	96.9	98.0
✓	×	✓	93.5	95.8	96.2
✓	✓	✓	**95.0**	**98.5**	**98.8**

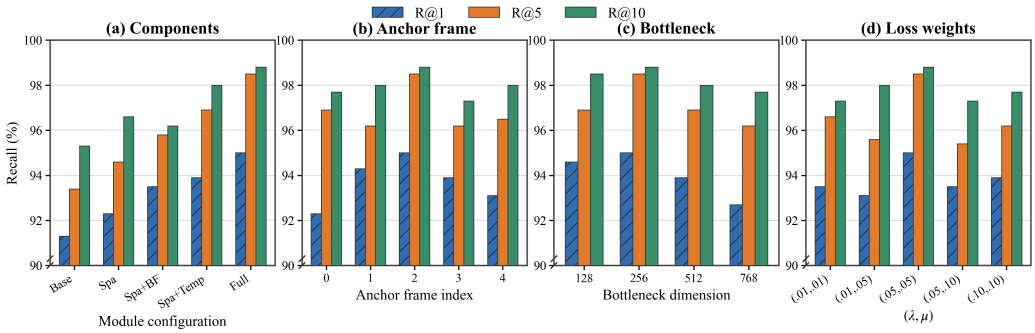


Figure 3: Ablation studies on EPRBench: (a) model components, (b) anchor-frame index, (c) bottleneck dimension, and (d) alignment-loss weights.

A bottleneck dimension of 256, balanced loss weights $(\lambda, \mu) = (0.05, 0.05)$, and the middle anchor frame ($t = 2$) provide the best overall performance.

4.5 Visualization and Limitation Analysis

Figure 3 summarizes the ablation results of the proposed components and key design choices. Spatial OT improves the SALAD-based baseline from 91.3% to 92.3% in R@1, while integrating temporal OT and bottleneck fusion further increases R@1 to 95.0%. The middle anchor frame ($a = 2$), a bottleneck dimension of 256, and balanced alignment-loss weights $(\lambda, \mu) = (0.05, 0.05)$ achieve the best overall performance. Additional qualitative retrieval results are provided in Appendix G.

While MSTOT-EPR performs well on both event-based benchmarks, it relies on paired text descriptions and is therefore sensitive to annotation noise and domain shift. Moreover, iterative spatial-temporal OT increases computational cost, while the fixed five-frame window may be suboptimal under rapidly varying event rates. Future work will explore annotation-efficient semantic supervision, sparse OT solvers, and adaptive temporal aggregation.

5 Conclusion

In this work, we present MSTOT-EPR, an optimal transport-driven framework that bridges high-level text semantics with asynchronous event streams for robust visual place recognition. MSTOT-EPR realizes bidirectional cross-modal alignment via local OT and global MMD supervision with bottleneck fusion, while jointly modeling multi-scale spatial structures and inter-frame dynamics before SALAD-based aggregation. Extensive experiments on NYC-Event-VPR and EPRBench demonstrate our state-of-the-art retrieval accuracy on event data alongside competitive RGB generalization. Ablations confirm the symbiotic contributions of semantic, spatial, and temporal modules. Future work will focus on weakly-supervised semantic distillation, efficient transport solvers, and rate-adaptive temporal aggregation.

AI USE STATEMENT

In preparing this manuscript, we used generative AI tools (large language models) solely for language refinement and proofreading. These tools were not used to formulate research ideas, design or conduct experiments, implement code, generate figures, or analyze data. All AI-assisted content was carefully reviewed and revised by the authors, who take full responsibility for the accuracy and integrity of the final manuscript.

REFERENCES

Amar Ali-bey, Brahim Chaib-draa, and Philippe Giguère. Convap: Revisiting convolutional features for visual place recognition. In *WACV*, pp. 4702–4711, 2022.

Amar Ali-Bey, Brahim Chaib-Draa, and Philippe Giguere. Mixvpr: Feature mixing for visual place recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2998–3007, 2023.

Amar Ali-bey, Brahim Chaib-draa, and Philippe Giguère. BoQ: A place is worth a bag of learnable queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17794–17803, June 2024.

Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5297–5307, 2016.

Gabriele Berton, Carlo Masone, and Barbara Caputo. Rethinking visual geo-localization for large-scale applications. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4868–4878, 2022a. doi: 10.1109/CVPR52688.2022.00483.

Gabriele Berton, Riccardo Mereu, Gabriele Trivigno, Carlo Masone, Gabriela Csurka, Torsten Sattler, and Barbara Caputo. Deep visual geo-localization benchmark. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5386–5397, 2022b. doi: 10.1109/CVPR52688.2022.00532.

Gabriele Berton, Gabriele Trivigno, Barbara Caputo, and Carlo Masone. Eigenplaces: Training viewpoint robust models for visual place recognition. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11046–11056, 2023. doi: 10.1109/ICCV51070.2023.01017.

Yin Bi, Aaron Chadha, Alhabib Abbe, Yonina Eldar, and Miguel R. D. Rodrigues. Graph-based asynchronous event processing for machine vision applications. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 311–320, 2019.

Lan Chen, Yu Jin, Xiao Wang, Bo Jiang, Lin Zhu, Yonghong Tian, Jin Tang, and Bin Luo. Dynamic graph induced contour-aware heat conduction network for event stream based object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2026a.

Shunpeng Chen, Changwei Wang, Rongtao Xu, Jinzhou Lin, Wenhao Xu, Li Guo, Shibiao Xu, et al. Sage: Spatial-visual adaptive graph exploration for efficient visual place recognition. In *International conference on learning representations*, volume 2026, pp. 28660–28683, 2026b.

Nicolas et al. Courty. Optimal transport for domain adaptation. *IEEE TPAMI*, 2017.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

Tianchen Deng, Xun Chen, Ziming Li, Hongming Shen, Danwei Wang, Javier Civera, and Hesheng Wang. Unipr-3d: Towards universal visual place recognition with 3d visual geometry grounded transformer. In *Arxiv*, 2026.

Tobias Fischer and Michael Milford. Eventvlad: Visual place recognition with event cameras. *IEEE Robotics and Automation Letters*, 5(2):2558–2565, 2020.

Tobias Fischer and Michael Milford. How many events do you need? event-based visual place recognition using sparse but varying pixels. *IEEE Robotics and Automation Letters*, 7(4):12275–12282, 2022. doi: 10.1109/LRA.2022.3216226.

Charlie et al. Frogner. Learning with a wasserstein loss. In *NeurIPS*, 2015.

Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J Davison, Jörg Conradt, Kostas Daniilidis, et al. Event-based vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(1):154–180, 2020a.

Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Cesi, Stefan Leutenegger, Andrew J. Davison, Davide Scaramuzza, and Ryad Benosman. Event-based vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):154–180, 2020b.

Stephen Hausler and Peyman Moghadam. Pair-vpr: Place-aware pre-training and contrastive pair classification for visual place recognition with vision transformers. *IEEE Robotics and Automation Letters*, 10(4):4013–4020, 2025. doi: 10.1109/LRA.2025.3546512.

Adam D. Hines, Gokul B. Nair, Nicolás Marticorena, Michael Milford, and Tobias Fischer. Event-gem: Global-to-local feature matching for event-based visual place recognition, 2026. URL `https://arxiv.org/abs/2603.05807`.

Nikhil Keetha, Avneesh Mishra, Jay Karhade, Piyush Gopalakrishnan, Sebastian Scherer, Sanjiv Singh, and Sourav Garg. Anyloc: Towards universal visual place recognition. In *IEEE Robotics and Automation Letters*, volume 8, pp. 7462–7469, 2023.

Nicholas et al. Kolkin. Style transfer by relaxed optimal transport and self-similarity. In *CVPR*, 2019.

Delei Kong, Zheng Fang, Kuanxu Hou, Haojia Li, Junjie Jiang, Sonya Coleman, and Dermot Kerr. Event-vpr: End-to-end weakly supervised deep network architecture for visual place recognition using event-based vision sensor. *IEEE Transactions on Instrumentation and Measurement*, 71: 1–18, 2022.

Junho Lee, Eung-Chang Kim, and Ayoung Kim. Event-based visual place recognition: A survey and performance evaluation. *IEEE Transactions on Robotics*, 39(5):3812–3829, 2023.

Zhenyu Li and Tianyi Shang. A^2GC: Asymmetric aggregation with geometric constraints for locally aggregated descriptors. *arXiv preprint arXiv:2511.14109*, 2025.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding and mitigating domain gaps in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18608–18617, 2022.

Feng Lu, Xiangyuan Lan, Lijun Zhang, Dongmei Jiang, Yaowei Wang, and Chun Yuan. Cricavpr: Cross-image correlation-aware representation learning for visual place recognition. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16772–16782, 2024a. doi: 10.1109/CVPR52733.2024.01587.

Feng Lu, Lijun Zhang, Xiangyuan Lan, Shuting Dong, Yaowei Wang, and Chun Yuan. Towards seamless adaptation of pre-trained models for visual place recognition. *arXiv preprint arXiv:2402.14505*, 2024b.

Feng Lu, Xinyao Zhang, Canming Ye, Shuting Dong, Lijun Zhang, Xiangyuan Lan, and Chun Yuan. Supervlad: Compact and robust image descriptors for visual place recognition. In *Advances in Neural Information Processing Systems*, volume 37, pp. 5789–5816, 2024c.

Feng Lu, Tong Jin, Xiangyuan Lan, Lijun Zhang, Yunpeng Liu, Yaowei Wang, and Chun Yuan. Selavpr++: Towards seamless adaptation of foundation models for efficient place recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025a.

Feng Lu, Tong Jin, Canming Ye, Xiangyuan Lan, Yunpeng Liu, and Chun Yuan. Towards implicit aggregation: Robust image representation for place recognition in the transformer era. In *The Annual Conference on Neural Information Processing Systems*, 2025b.

Reza Mahjoub and Shohreh Kasaei. Salad: Self-supervised aggregation learning for adverse conditions in visual place recognition. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 11518–11524, 2023. doi: 10.1109/ICRA48891.2023.10160968.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research Journal*, pp. 1–31, 2024.

Taiyi Pan, Junyang He, Chao Chen, Yiming Li, and Chen Feng. Nyc-event-vpr: A large-scale high-resolution event-based visual place recognition dataset in dense urban environments. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4657–4664. IEEE, 2025.

Gabriel Peyré and Marco Cuturi. *Computational Optimal Transport*. Now Publishers, 2019.

Qibo Qiu, Shun Zhang, Haiming Gao, Honghui Yang, Haochao Ying, Wenxiao Wang, and Xiaofei He. Emvp: Embracing visual foundation model for visual place recognition with centroid-free probing. *Advances in Neural Information Processing Systems*, 37:120928–120950, 2024.

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

Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. Events-to-video: Bringing modern computer vision to event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3857–3866, 2019.

Paul-Edouard et al. Sarlin. Superglue: Learning feature matching with graph neural networks. In *CVPR*, 2020.

Bernhard Schmitzer. Stabilized sparse scaling algorithms for entropy regularized transport problems. *SIAM Journal on Scientific Computing*, 41(3):A1443–A1481, 2019.

Cédric Villani. *Optimal Transport: Old and New*. Springer, 2009.

Changwei Wang, Shunpeng Chen, Yukun Song, Rongtao Xu, Zherui Zhang, Jiguang Zhang, Haoran Yang, Yu Zhang, Kexue Fu, Shide Du, et al. Focus on local: Finding reliable discriminative regions for visual place recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7536–7544, 2025.

Ruotong Wang, Yipeng Shen, Weidong Zuo, Sanping Zhou, and Nanning Zheng. Transvpr: Transformer-based visual place recognition with feature alignment for challenging scenes. In *ICRA*, pp. 8350–8356, 2022.

Xiao Wang, Haiyang Wang, Shiao Wang, Qiang Chen, Jiandong Jin, Haoyu Song, Bo Jiang, and Chenglong Li. Rgb-event based pedestrian attribute recognition: A benchmark dataset and an asymmetric rwkv fusion framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 32745–32755, 2026a.

Xiao Wang, Xingxing Xiong, Jinfeng Gao, Xufeng Lou, Bo Jiang, Sibao Chen, Yaowei Wang, and Yonghong Tian. Eprbench: A high-quality benchmark dataset for event stream based visual place recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 2026b.

Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R Scott. Multi-similarity loss with general pair weighting for deep metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5022–5030, 2019.

Jun Xu, Qian Liu, Wenbing Huang, Jian Wang, and Zhiguo Li. R2former: Robust and reliable transformer for visual place recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12251–12260, 2023.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.

Weiliang Zuo, Liguo Liu, Yizhe Li, Yanqing Shen, Fuhua Xiang, Jingmin Xin, and Nanning Zheng. Prgs: Patch-to-region graph search for visual place recognition. *Pattern Recognition*, pp. 111673, 2025.

A Background and Applications of Optimal Transport

Optimal Transport (OT) provides a principled framework for comparing probability distributions by finding a transport plan that minimizes the cost of moving probability mass between them Villani (2009); Peyré & Cuturi (2019). Unlike pointwise metrics or global pooling, OT captures soft many-to-many correspondences under marginal constraints. Entropic regularization enables efficient and differentiable optimization through Sinkhorn iterations Cuturi (2013), making OT practical for modern representation learning.

OT has been applied to domain adaptation Courty (2017), distribution-aware learning objectives Frogner (2015), style transfer Kolkin (2019), and feature matching Sarlin (2020). In visual place recognition, SALAD Mahjoub & Kasaei (2023) employs OT-inspired assignments to aggregate local image features into compact global descriptors. However, existing OT-based VPR methods mainly focus on static RGB images and use OT for spatial matching or feature aggregation.

Event-based VPR additionally requires heterogeneous event-text alignment and the modeling of spatial and temporal event structures. Unlike prior methods that apply OT within a single modality or representation level, our approach introduces OT along three dimensions: event-text token alignment, multi-scale spatial alignment, and inter-frame temporal alignment. It therefore extends OT from static feature aggregation to joint cross-modal, cross-scale, and temporal modeling for text-assisted event-stream VPR.

B Log-Domain Sinkhorn Solver

All optimal-transport modules in MSTOT-EPR are solved using the same log-domain Sinkhorn procedure. Given a cost matrix $\mathbf{C} \in \mathbb{R}^{n\times m}$, source marginal $\boldsymbol{\mu} \in \mathbb{R}^n$, target marginal $\boldsymbol{\nu} \in \mathbb{R}^m$, and entropic regularization coefficient ϵ, we define $\beta = -1/\epsilon$.

Let $\mathbf{f}^{(k)} \in \mathbb{R}^n$ and $\mathbf{g}^{(k)} \in \mathbb{R}^m$ be the log-domain scaling vectors at iteration k. We initialize $\mathbf{f}^{(0)} = \mathbf{0}$ and $\mathbf{g}^{(0)} = \mathbf{0}$, and update them as

$$f_i^{(k+1)} = \log \mu_i - \log \sum_{j=1}^{m} \exp\left(\beta C_{ij} + g_j^{(k)}\right), \tag{21}$$

$$g_j^{(k+1)} = \log \nu_j - \log \sum_{i=1}^{n} \exp\left(\beta C_{ij} + f_i^{(k+1)}\right). \tag{22}$$

After K iterations, the transport plan is recovered as

$$P_{ij} = \exp\left(\beta C_{ij} + f_i^{(K)} + g_j^{(K)}\right). \tag{23}$$

These alternating updates enforce the prescribed row and column marginals. Performing the computation in the log domain improves numerical stability. We use this solver for cross-modal, multi-scale spatial, and inter-frame temporal optimal transport.

C Detailed Training Objectives

C.1 Retrieval Loss

Given a mini-batch of N normalized place descriptors $\{\mathbf{d}_i\}_{i=1}^{N}$, their pairwise similarity is $s_{ij} = \mathbf{d}_i^\top \mathbf{d}_j$. For each anchor i, the Multi-Similarity miner selects informative positive and negative sets $\mathcal{S}_i^+$ and $\mathcal{S}_i^-$. The retrieval loss is

$$\mathcal{L}_{\text{retr}} = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{\alpha}\log\left(1+\sum_{p\in\mathcal{S}_i^+} e^{-\alpha s_{ip}-\tau}\right)+\frac{1}{\beta}\log\left(1+\sum_{n\in\mathcal{S}_i^-} e^{\beta s_{in}-\tau}\right)\right]. \tag{24}$$

We use $\alpha = 1$, $\beta = 50$, $\tau = 0$, and a mining margin of $\epsilon_m = 0.1$.

C.2 Local Optimal-Transport Alignment

Using the cross-modal cost matrix $\mathbf{C}_{i,a}^{\text{loc}}$ and transport plan $\mathbf{P}_{i,a}^{\text{loc}}$ defined in Sec. 3.3, the loss evaluated on anchor frame a is

$$\mathcal{L}_{\text{loc}} = \frac{1}{N}\sum_{i=1}^{N}\left\langle \mathbf{P}_{i,a}^{\text{loc}}, \mathbf{C}_{i,a}^{\text{loc}}\right\rangle_F. \tag{25}$$

Gradients are not propagated through the transport plan during optimization.

C.3 Global Distribution Alignment

Let $\bar{\mathbf{c}}_i$ and $\bar{\mathbf{s}}_i$ denote the CLS feature of the projected and normalized event and the sentence feature, respectively. Their distribution discrepancy is minimized using

$$\mathcal{L}_{\text{mmd}} = \frac{1}{N^2}\sum_{i,j=1}^{N}\left[k(\bar{\mathbf{c}}_i,\bar{\mathbf{c}}_j)+k(\bar{\mathbf{s}}_i,\bar{\mathbf{s}}_j)-2k(\bar{\mathbf{c}}_i,\bar{\mathbf{s}}_j)\right]. \tag{26}$$

Here, $k(\cdot,\cdot)$ is a multi-kernel Gaussian kernel composed of five geometrically scaled bandwidths.

D Implementation Details

We implement MSTOT-EPR using PyTorch Lightning and train it on a single NVIDIA RTX 3090 GPU. Each input sample consists of five consecutive event frames resized to 224×224. The frames are normalized using ImageNet statistics and augmented with RandAugment during training. Each mini-batch contains 20 places, with four event sequences sampled for each place. We use pretrained DINOv2 ViT-B/14 as the visual backbone and fine-tune only its last four Transformer blocks, while keeping the CLIP ViT-L/14 text encoder frozen. The spatial OT module constructs multi-scale representations using scale factors of 0.5, 1.0, and 2.0. The temporal OT module aligns all five frames to the middle frame ($a = 2$). All OT modules perform 20 log-domain Sinkhorn iterations with an entropic regularization coefficient of 0.08. The bottleneck dimension is set to 256, and the alignment-loss weights are set to $(\lambda, \mu) = (0.05, 0.05)$. SALAD uses 64 clusters and produces an 8448-dimensional ℓ_2-normalized place descriptor. The model is trained using the Multi-Similarity loss and Multi-Similarity Miner. We use AdamW with an initial learning rate of 6×10^{-5}, a weight decay of 9.5×10^{-9}, and a linear learning-rate scheduler. Training lasts for 20 epochs using mixed-precision computation.

E Detailed Comparison with State-of-the-Art Methods

NYC-Event-VPR-Event. MSTOT-EPR achieves 65.0%, 78.9%, and 80.9% in R@1, R@5, and R@10, respectively. Compared with the best competing result for each metric, the corresponding improvements are 1.0, 3.7, and 2.0 percentage points. The improvements indicate that multi-scale spatial modeling and inter-frame alignment facilitate the extraction of discriminative place representations from sparse event observations.

Table 3: Analysis of bottleneck fusion dimension.

Dimension	R@1	R@5	R@10
128	94.6	96.9	98.5
256	**95.0**	**98.5**	**98.8**
512	93.9	96.9	98.0
768	92.7	96.2	97.7

Table 4: Alignment-loss weight analysis.

Weight (λ, μ)	R@1	R@5	R@10
$(0.01, 0.01)$	93.5	96.6	97.3
$(0.01, 0.05)$	93.1	95.6	98.0
$(\mathbf{0.05}, \mathbf{0.05})$	**95.0**	**98.5**	**98.8**
$(0.05, 0.1)$	93.5	95.4	97.3
$(0.1, 0.1)$	93.9	96.2	97.7

Table 5: Anchor-frame selection analysis.

Index t	R@1	R@5	R@10
0	92.3	96.9	97.7
1	94.3	96.2	98.0
2	**95.0**	**98.5**	**98.8**
3	93.9	96.2	97.3
4	93.1	96.5	98.0

EPRBench. MSTOT-EPR obtains 95.0%, 98.5%, and 98.8% in R@1, R@5, and R@10, outperforming the best competing results by 0.7, 2.4, and 1.7 percentage points, respectively. The consistent improvements suggest that textual semantics complement sparse event observations, while the spatial-temporal OT modules preserve structural and temporal consistency under challenging environmental variations.

NYC-Event-VPR-RGB. On the RGB split, MSTOT-EPR achieves 81.5%, 89.7%, and 90.7% in R@1, R@5, and R@10, respectively. Although it does not outperform RGB-specific approaches, the results show that the learned representation retains some generalization to conventional visual inputs. This performance difference is expected because MSTOT-EPR is specifically designed to address the sparsity and asynchronous dynamics of event data.

F Additional Ablation Studies

Bottleneck dimension. Table 3 evaluates the bottleneck dimension. The model achieves the best performance with a dimension of 256, reaching 95.0% R@1, 98.5% R@5, and 98.8% R@10. Increasing the dimension to 512 or 768 reduces R@1 to 93.9% and 92.7%, respectively, suggesting that an excessively large bottleneck introduces redundant parameters and increases the risk of overfitting.

Alignment-loss weights. Table 4 studies the weights of the local OT and global MMD losses. The balanced setting $(\lambda, \mu) = (0.05, 0.05)$ achieves the best overall performance. Smaller weights provide insufficient cross-modal supervision, whereas larger weights may overwhelm the primary retrieval objective.

Anchor-frame selection. Table 5 compares different temporal anchors. The middle frame ($t = 2$) performs best because it receives contextual information from both preceding and subsequent frames.

G Additional Qualitative Results

Figure 4 presents qualitative top-1 retrieval comparisons on EPRBench. The selected queries cover diverse and challenging conditions, including high-rise buildings with repetitive facade patterns, oblique observations of similar architectural structures, vegetation-dominated scenes, and road environments containing limited structural details. In these cases, event observations mainly preserve intensity-change contours while providing limited texture and appearance information. Consequently, different places may exhibit similar event patterns, whereas observations of the same place can vary considerably under changes in viewpoint, motion, and event density. As shown in Figure 4, competing methods correctly retrieve several relatively distinctive scenes but produce inconsistent



Figure 4: Qualitative top-1 retrieval comparison on EPRBench. Each row shows an event-stream query and the corresponding top-1 retrieved reference image from MSTOT-EPR (ours) and competing methods. Green and red borders denote correct and incorrect retrievals, respectively. Our method retrieves the correct places more consistently under viewpoint changes, sparse event observations, and challenging scene appearances.

results when local structures are repetitive or only partially observed. In contrast, MSTOT-EPR returns the correct reference for all the presented queries, indicating more stable retrieval behavior across different scene types.

The qualitative results are also consistent with the design of MSTOT-EPR. First, the event-text alignment module establishes fine-grained correspondences between event patches and textual tokens, providing complementary scene-level semantics when event appearance alone is ambiguous. Second, the multi-scale spatial OT module aligns event representations using both feature similarity and spatial position, allowing local contours and global layouts to be compared across different resolutions. This is particularly relevant to scenes involving repetitive facades, scale changes, or partial structural observations. Third, the inter-frame temporal OT module transports neighboring frame tokens onto a shared anchor support and adaptively aggregates the aligned features. It therefore exploits complementary observations distributed across consecutive event frames instead of relying on a single sparse representation. The improved retrieval consistency shown in these examples qualitatively supports the joint use of semantic guidance, cross-scale spatial correspondence, and inter-frame temporal alignment for event-based place recognition.