International Journal of Creative and Open Research in Engineering and Management
ISSN: **3108-1754 (Online)**
Volume 02 Issue 05 May-2026 | Impact Factor: 3.5

Representational Divergence Between Spiking and Non-Spiking Neural Architectures Under Multimodal Contrastive Learning

Mrs. G. Archana [1], Kumaran V[2] Amarjith M[3]
[1]Assistant Professor, Department of AIDS,
Dhanalakshmi Srinivasan University, Trichy, India
[2] Department of Artificial Intelligence and Data Science,
Dhanalakshmi Srinivasan University, Trichy, India
[3] Department of Computer Science Engineering,
Dhanalakshmi Srinivasan University, Trichy, India
Emails: archana@dsuniversity.ac.in kumaran.v.4002@gmail.com amarjithm26@gmail.com

Abstract—Multimodal representation learning enables im-age–text alignment through shared embedding spaces optimized using contrastive objectives. Although Spiking Neural Networks (SNNs) provide biologically inspired temporal computation, their representational behavior under static multimodal supervision remains insufficiently explored. This study presents a controlled comparison between a Spiking Neural Network (SNN) and a Multilayer Perceptron (MLP) for multimodal image–text re-trieval. To isolate the effect of temporal spiking dynamics, both architectures were trained under identical embedding dimension-ality, optimization settings, contrastive learning objectives, and retrieval protocols. Experiments were conducted on a balanced episodic benchmark consisting of 1,000 image–text pairs. Results show that both architectures learn stable and non-collapsed embedding spaces with broad cosine similarity distributions. However, retrieval performance for both models remains close to chance under static supervision. Despite comparable task-level performance, cross-model cosine similarity analysis reveals substantial representational divergence between SNN and MLP embeddings, indicating distinct embedding geometries under identical learning conditions. The findings suggest that tempo-ral spiking dynamics alone do not improve static multimodal retrieval alignment and highlight a mismatch between spiking inductive bias and temporally unstructured supervision. Overall, the study emphasizes the importance of representation-level anal-ysis alongside conventional retrieval evaluation for neuromorphic multimodal learning systems.

Index Terms—Spiking Neural Networks, Multimodal Retrieval, Contrastive Learning, Neuromorphic Computing, Embedding Geometry, Representation Learning

I. Introduction

Multimodal representation learning has become an important research direction in artificial intelligence due to its ability to integrate heterogeneous modalities such as images and text into a shared semantic embedding space. These representations support applications including cross-modal retrieval, semantic search, visual grounding, and multimodal reasoning. The primary objective is to align semantically related image–text pairs while maintaining separation between unrelated samples within the embedding space.

Recent advances in multimodal learning have been driven by contrastive learning frameworks such as InfoNCE and normalized temperature-scaled cross-entropy (NT-Xent). These objectives maximize similarity between matched multimodal pairs and minimize similarity between unrelated samples. Such methods have demonstrated strong performance in retrieval and representation learning tasks, while also motivating research on embedding geometry, representation alignment, and latent space organization.

Despite these developments, most multimodal contrastive systems rely on conventional artificial neural networks (ANNs) with continuous-valued activations. Comparatively fewer studies have explored biologically inspired Spiking Neural Networks (SNNs) under equivalent multimodal learning objectives. Unlike traditional ANNs, SNNs process information through temporally discrete spike events using membrane integration and event-driven computation, making them attractive for neuromorphic and energy-efficient computing.

However, current image–text retrieval benchmarks are predominantly static and contain little temporal structure. Consequently, it remains unclear whether temporal spiking dynamics meaningfully influence embedding formation, representation geometry, or retrieval behavior under static multimodal supervision.

Previous comparisons between spiking and non-spiking architectures have mainly focused on classification accuracy, energy efficiency, or ANN-to-SNN conversion methods. Controlled representational comparisons between SNNs and conventional Multilayer Perceptrons (MLPs) under identical multimodal contrastive learning settings remain limited. Existing studies often differ in optimization strategies, architectural capacity, or dataset assumptions, making it difficult to isolate the contribution of temporal spiking dynamics.

Motivated by these limitations, this work presents a controlled comparative study between SNNs and MLPs for multimodal image–text retrieval. Both architectures are trained under identical embedding dimensionality, optimization set-

tings, contrastive objectives, and retrieval protocols to isolate the influence of temporal spiking computation. Rather than emphasizing state-of-the-art retrieval performance, the study focuses on diagnostic representational analysis, including embedding stability, cosine similarity distributions, cross-model alignment, and latent embedding geometry.

The main contributions of this work are summarized as follows:

1) A controlled comparison between Spiking Neural Networks (SNNs) and Multilayer Perceptrons (MLPs) under identical multimodal contrastive learning conditions.
2) A diagnostic analysis of embedding geometry, cosine similarity structure, and cross-model representational alignment.
3) Experimental evidence showing that spiking and non-spiking architectures can exhibit comparable retrieval performance while producing substantially different embedding geometries.
4) An interpretative analysis of the mismatch between temporal spiking inductive bias and static multimodal supervision.

II. Related Work

This section reviews prior research on multimodal contrastive learning, spiking neural networks (SNNs), and representation geometry analysis. The discussion focuses on controlled comparisons between spiking and non-spiking architectures under multimodal supervision.

A. Contrastive Multimodal Learning

Contrastive learning has become a dominant approach for multimodal representation learning, particularly in image–text retrieval tasks. Objectives such as InfoNCE and normalized temperature-scaled cross-entropy (NT-Xent) align semantically related image–text pairs while separating unrelated samples in a shared embedding space. Recent studies have shown strong retrieval capability using contrastive supervision combined with pretrained encoders.

Beyond retrieval accuracy, recent work has emphasized embedding geometry, alignment, and similarity structure within latent spaces. However, most existing multimodal systems rely on conventional artificial neural networks (ANNs), while the effect of alternative neural computation paradigms on representation geometry remains less explored. Moreover, current image–text benchmarks are largely static and contain little temporal structure.

B. Spiking Neural Networks

Spiking Neural Networks (SNNs) are biologically inspired models that process information using discrete spike events and temporal membrane dynamics. Compared with conventional ANNs, SNNs employ sparse event-driven communication and temporal integration, making them suitable for neuromorphic and energy-efficient computation.

Recent advances in surrogate gradient learning have enabled effective optimization of deep SNNs. Prior work has demonstrated strong performance in event-driven vision, auditory processing, and temporal sensory applications. However, most SNN research focuses on biological plausibility, hardware efficiency, or temporal signal processing rather than multimodal representation learning.

C. Spiking Models for Multimodal Learning

Research on multimodal spiking architectures remains relatively limited. Existing studies mainly investigate event-based sensory fusion, neuromorphic robotics, or associative memory systems. In many cases, temporal encoding schemes or sequential structures are explicitly introduced to support spike-based computation.

Direct comparisons between SNNs and non-spiking baselines are often complicated by differences in architecture, optimization, and preprocessing pipelines. Furthermore, relatively few studies analyze embedding geometry or cross-modal representation alignment under controlled experimental settings.

D. Representation Geometry Analysis

Representation geometry has become increasingly important in understanding learned embedding spaces. Techniques such as cosine similarity analysis, principal component analysis (PCA), and t-SNE visualization have shown that models with similar task-level performance may nevertheless learn substantially different latent organizations.

Most existing analyses focus primarily on ANN-based architectures. Comparatively fewer studies investigate how temporal spiking dynamics influence embedding structure under identical supervision conditions. This gap is particularly important in multimodal contrastive learning, where embedding geometry directly affects retrieval behavior.

E. Positioning of the Present Work

The present work adopts a controlled representational analysis framework for comparing SNNs and multilayer perceptrons (MLPs) in multimodal image–text retrieval. Unlike prior studies emphasizing benchmark performance, this study isolates temporal spiking dynamics as the principal experimental variable while maintaining identical embedding dimensionality, optimization settings, and contrastive objectives across architectures.

In addition to retrieval evaluation, the work investigates embedding stability, cosine similarity structure, and cross-model representational divergence. The study therefore provides empirical insight into how architectural inductive bias influences multimodal representation geometry under static contrastive supervision.

III. Methodology

This section describes the controlled experimental framework used to compare spiking and non-spiking neural representations for multimodal retrieval. The methodology is specifically designed to isolate the influence of temporal spiking dynamics while minimizing the impact of confounding factors such as optimization strategy, architectural capacity, and

data exposure. To ensure reproducibility and interpretability, identical training procedures, hyperparameters, and evaluation protocols are applied across all experimental conditions.

A. Experimental Design

The primary objective of this study is to investigate the effect of temporal spiking dynamics on multimodal representation learning under static contrastive supervision. For controlled comparison, the Spiking Neural Network (SNN) and Multilayer Perceptron (MLP) are configured with identical embedding dimensionality, hidden layer structure, optimization settings, and training schedules. The principal distinction between the two architectures is the presence or absence of temporal spiking computation.

The experimental framework is designed primarily for representational analysis rather than maximization of retrieval accuracy. Under this controlled setting, observed differences in embedding geometry or retrieval behavior can be interpreted as consequences of the underlying neural computation mechanism rather than disparities in optimization capacity or architectural scale.

B. Architectural Parity

Both architectures process multimodal inputs and project them into a shared 128-dimensional embedding space. The image modality is represented using 1024-dimensional feature vectors, while text inputs are represented using 64-dimensional embeddings.

To ensure strict architectural parity, both models employ identical hidden-layer dimensionality, projection dimensionality, dropout configuration, and output normalization strategy. No architecture-specific loss functions or additional tuning procedures are introduced for either model. This design minimizes confounding effects and enables representational differences to be attributed primarily to temporal spiking dynamics.

Fig. 1. Architecture comparison between the spiking neural network (SNN) encoder and the multilayer perceptron (MLP) baseline under controlled multimodal contrastive learning conditions.

TABLE I
ARCHITECTURE COMPARISON BETWEEN THE SPIKING NEURAL NETWORK (SNN) ENCODER AND THE MULTILAYER PERCEPTRON (MLP) BASELINE UNDER CONTROLLED EXPERIMENTAL PARITY.

Component	SNN Encoder	MLP Encoder
Image Input Dim	1024	1024
Text Input Dim	64	64
Hidden Layer 1	256 LIF	256 FC + ReLU
Hidden Layer 2	128 LIF	128 FC + ReLU
Timesteps	60	None
Embedding Dim	512	512
Dropout	0.02	0.02

C. Spiking Neural Network Encoder

The spiking encoder is implemented using a feed-forward architecture composed of Leaky Integrate-and-Fire (LIF) neurons. Continuous-valued input features are transformed into temporal spike trains through rate-based encoding over a fixed simulation window. Spike activity propagates through successive spiking layers and is aggregated temporally to produce static embedding representations suitable for multimodal contrastive learning.

The membrane dynamics follow the standard LIF formulation:

$$\tau_m \frac{dV(t)}{dt} = -V(t) + I(t) \quad (1)$$

where $V(t)$ denotes the membrane potential, $I(t)$ represents the synaptic input current, and τ_m is the membrane time constant. A spike is emitted whenever the membrane potential exceeds a predefined threshold, after which the membrane state is reset.

Since the spiking activation function is non-differentiable, surrogate gradient approximations are employed during backpropagation to enable gradient-based optimization.

D. Multilayer Perceptron Baseline

The non-spiking baseline consists of a conventional Multilayer Perceptron (MLP) composed of fully connected layers and nonlinear activation functions. The MLP maintains the same hidden dimensionality, network depth, and embedding dimensionality as the SNN but performs static feed-forward computation without temporal expansion or spike generation.

This architecture serves as a controlled non-temporal baseline, allowing temporal spiking dynamics to remain the principal independent variable within the experimental comparison.

E. Spike Encoding Configuration

Static input features are converted into spike trains using rate-based Poisson encoding. The spike probability is proportional to the normalized input magnitude over a fixed simulation horizon of 60 timesteps. This encoding strategy introduces temporal activity without imposing task-specific sequential assumptions.

Rate-based encoding preserves semantic feature content while enabling temporally distributed neural computation. Importantly, no artificial temporal ordering or event-driven

structure is added to the dataset, ensuring that the analysis focuses specifically on the interaction between temporal spiking dynamics and static multimodal supervision.

TABLE II
SPIKE ENCODING CONFIGURATION AND TEMPORAL PROCESSING PARAMETERS USED FOR THE SPIKING NEURAL NETWORK.

Parameter	Value
Encoding Method	Rate-based
Timesteps	60
Membrane Constant	12.0
Threshold	0.9
Image Sparsity	0.9368
Caption Sparsity	0.8359

F. Contrastive Learning Objective

Both architectures are optimized using the normalized temperature-scaled cross-entropy (NT-Xent) loss. The objective encourages embeddings of semantically related image–text pairs to become similar while simultaneously separating unrelated samples within the batch.

Cosine similarity between embeddings is defined as:

$$\text{sim}(x, y) = \frac{x \cdot y}{|x|\,|y|} \tag{2}$$

The NT-Xent loss for a positive pair is formulated as:

$$L_i = -\log \sum_{k=1}^{N} \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\mathbf{1}_{[k \neq i]}\ \exp(\text{sim}(z_i, z_k)/\tau)} \tag{3}$$

where τ denotes the temperature parameter controlling similarity scaling. The same temperature value is maintained across all experiments to ensure optimization consistency.

G. Optimization Protocol

All models are trained using the Adam optimizer with identical hyperparameter configurations. Learning rate, batch size, weight decay, and training duration remain constant across architectures. No architecture-specific regularization, scheduling strategy, or early stopping mechanism is introduced.

Training is conducted for a fixed number of epochs to ensure equivalent exposure to the training data. Random seeds are fixed to reduce stochastic variability between experimental runs. The optimization procedure is intentionally conservative and standardized to minimize architecture-specific tuning effects.

TABLE III
SHARED OPTIMIZATION AND TRAINING HYPERPARAMETERS APPLIED UNIFORMLY ACROSS ALL EXPERIMENTAL CONDITIONS.

Parameter	Value
Optimizer	Adam
Epochs	24
Batch Size	16
Learning Rate	0.0003
Weight Decay	0.0001
Temperature	0.2

H. Embedding Normalization

All embeddings are ℓ_2 normalized prior to similarity computation:

$$\hat{z} = \frac{z}{|z|_2} \tag{4}$$

Normalization constrains representations to the unit hypersphere, stabilizing cosine similarity computations during both training and retrieval. This procedure also facilitates direct comparison of embedding geometry across architectures.

I. Retrieval Pipeline

Following training, embeddings from the test split are indexed using cosine similarity-based nearest-neighbor retrieval. Image and text representations are projected into the shared embedding space and evaluated through bidirectional cross-modal retrieval.

For each query embedding, cosine similarity is computed against all candidate embeddings from the opposite modality. Retrieval rankings are then generated according to similarity scores. The indexing and retrieval procedures remain identical across SNN, MLP, and random baseline conditions.

This unified retrieval framework ensures that observed differences in retrieval behavior arise from representational characteristics rather than variations in indexing or search procedures.

Fig. 2. Overview of the multimodal retrieval pipeline used for embedding generation, contrastive learning, and cosine similarity-based cross-modal retrieval evaluation.

J. Evaluation Setup

Evaluation encompasses both retrieval effectiveness and representational analysis. Retrieval performance is measured using Recall@1, Recall@5, and Mean Reciprocal Rank (MRR). These metrics assess the ability of the learned embedding space to preserve image–text correspondence.

In addition to rank-based evaluation, multiple embedding diagnostics are performed to examine representation stability and geometric organization. These analyses include embedding norm statistics, intra-model cosine similarity distributions, cross-model similarity analysis, and low-dimensional visualization using t-SNE and Principal Component Analysis (PCA).

The overall evaluation framework is intended not only to assess task-level performance but also to characterize the internal representational behavior produced by temporal and non-temporal neural computation under identical multimodal supervision conditions.

IV. Experimental Setup

A. Dataset and Benchmark Construction

Experiments were conducted using a controlled multimodal benchmark containing 1,000 image–text pairs. The dataset was constructed through a balanced episodic selection process designed to reduce semantic duplication and category imbalance. Samples were distributed across multiple episodic categories including human actions, outdoor scenes, animals, indoor environments, social interactions, emotional contexts, unusual events, and object-centric scenes. Each image was associated with a single representative caption.

A fixed train–test split was maintained throughout all experiments to ensure reproducibility and controlled comparison.

TABLE IV
STATISTICAL SUMMARY OF THE MULTIMODAL BENCHMARK USED FOR RETRIEVAL EVALUATION.

Property	Value
Categories	9
Caption Source	Flickr-style
Selection Strategy	Episodic-balanced
Final Benchmark Size	1000
Train Samples	800
Test Samples	200

B. Train–Test Partitioning

The dataset was divided using an 80/20 split consisting of 800 training pairs and 200 testing pairs. The same partition was used across all experiments for both the Spiking Neural Network (SNN) and Multilayer Perceptron (MLP) models.

No data augmentation, resampling, or curriculum scheduling was applied to preserve experimental consistency.

C. Feature Extraction and Input Preparation

Image features were extracted from a pretrained visual encoder, producing fixed 1024-dimensional embeddings. Text descriptions were encoded using a pretrained language model, generating 64-dimensional text embeddings.

All pretrained encoders remained frozen throughout training to isolate downstream representation learning. Prior to training, all features were standardized and normalized. Identical preprocessing was applied across architectures.

D. Training Configuration

Both architectures were trained using identical optimization settings. Training employed the Adam optimizer with learning rate 3×10^{-4}, batch size 16, and weight decay 1×10^{-4}. All models were trained for 24 epochs without early stopping.

The SNN and MLP shared identical embedding dimensionality, network depth, and optimization schedules. The primary experimental variable was the presence of temporal spiking dynamics.

Image–text alignment was learned using the NT-Xent objective defined in Eq. (3). All embeddings were ℓ_2 normalized before similarity computation. Random seeds were fixed to improve reproducibility.

E. Retrieval Evaluation Protocol

After training, multimodal retrieval was performed using cosine similarity-based nearest-neighbor matching. Retrieval was evaluated bidirectionally using both image-to-text and text-to-image queries.

A random retrieval baseline was additionally included to provide a lower-bound reference for performance comparison.

F. Evaluation Metrics

Retrieval performance was measured using Recall@1, Recall@5, and Mean Reciprocal Rank (MRR). Recall@K is defined as:

$$\text{Recall@K} = \frac{\text{Correct items in top-K}}{\text{Total queries}} \tag{5}$$

MRR is computed as:

$$\text{MRR} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\text{rank}_i} \tag{6}$$

In addition to retrieval accuracy, embedding quality was analyzed using cosine similarity statistics, embedding norm analysis, PCA, and t-SNE visualization to examine representation geometry and embedding diversity.

V. Results

This section presents the comparative evaluation of the Spiking Neural Network (SNN) and Multilayer Perceptron (MLP) for multimodal retrieval under identical supervision and optimization settings.

A. Retrieval Performance

Table V summarizes retrieval performance for the SNN, MLP, and random baseline. Both architectures achieve performance close to chance level across Recall@1, Recall@5, and Mean Reciprocal Rank (MRR). Although both models successfully optimize the contrastive objective, the learned embeddings fail to produce strong multimodal alignment.

The similar behavior of the SNN and MLP suggests that temporal spiking dynamics alone do not improve retrieval effectiveness under static multimodal supervision.

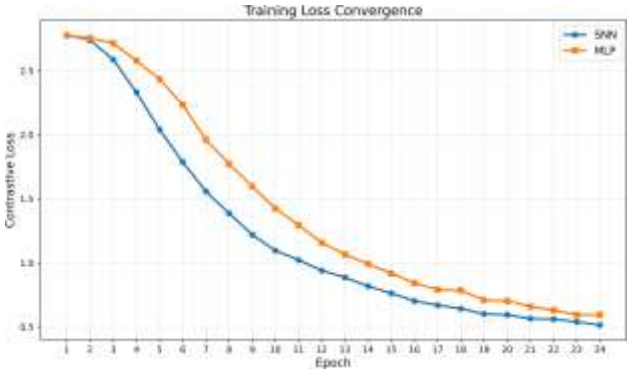

Fig. 3. Training convergence curves for the SNN and MLP under identical optimization settings.

TABLE V
MULTIMODAL RETRIEVAL PERFORMANCE ACROSS SNN, MLP, AND RANDOM BASELINE MODELS.

Model	Recall@1	Recall@5	MRR
SNN	0.000	0.005	0.002
MLP	0.001	0.004	0.003
Random	0.000	0.005	0.002

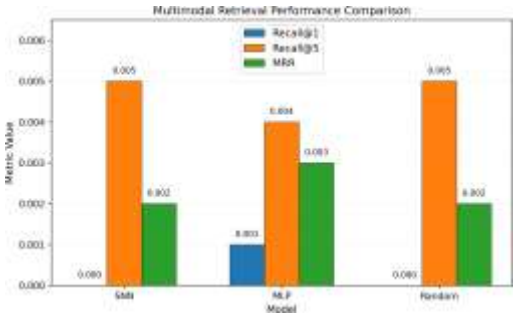

Fig. 4. Retrieval performance comparison across SNN, MLP, and random baseline models.

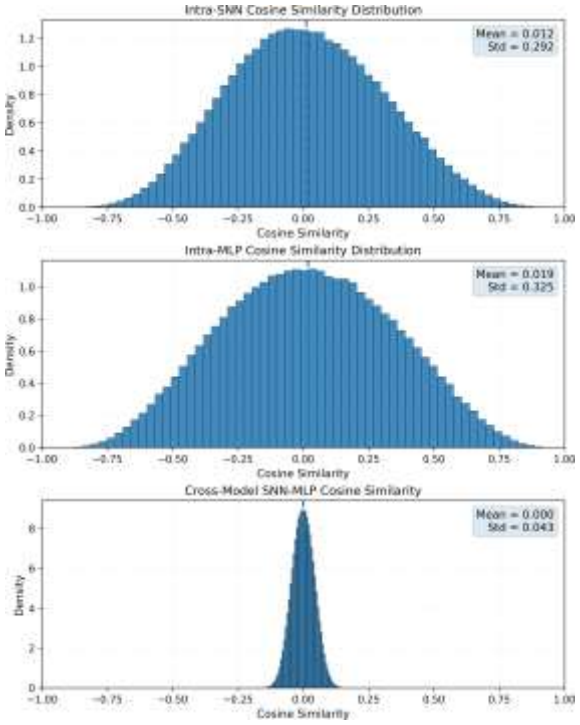

Fig. 5. Cosine similarity distributions for SNN, MLP, and cross-model embeddings.

B. Embedding Stability Analysis

Embedding norm statistics were analyzed to evaluate representation stability. As shown in Table VI, both architectures maintain normalized embeddings with stable cosine similarity distributions.

The results indicate that neither model suffers from representational collapse. Thus, poor retrieval performance is associated with weak cross-modal alignment rather than unstable embedding formation.

TABLE VI
EMBEDDING NORM AND COSINE SIMILARITY STATISTICS FOR LEARNED REPRESENTATIONS.

Metric	SNN	MLP
Norm Mean	1.0000	1.0000
Norm Std	0.0000	0.0000
Cosine Mean	0.0131	0.0199
Cosine Std	0.2937	0.3260

C. Cross-Model Representation Similarity

To compare representation geometry, cosine similarity was computed between SNN and MLP embeddings generated from identical inputs. The results in Table VII show near-zero mean similarity, indicating substantial representational divergence.

Despite identical training objectives and optimization settings, both architectures organize multimodal information differently within the embedding space. However, this divergence does not translate into improved retrieval performance.

TABLE VII
CROSS-MODEL COSINE SIMILARITY STATISTICS BETWEEN SNN AND MLP EMBEDDINGS.

Metric	Value
Mean SNN-MLP Cosine Similarity	-0.0015
Std SNN-MLP Cosine Similarity	0.0440

Fig. 6. PCA projection of SNN and MLP embeddings within the shared latent space.

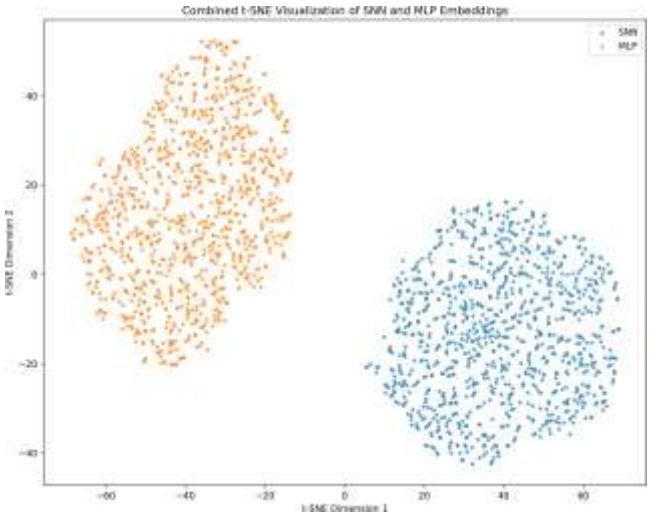

Fig. 7. t-SNE visualization of learned multimodal embeddings for the SNN and MLP architectures.

D. Representation Geometry

PCA and t-SNE visualizations further demonstrate that the SNN and MLP form distinct embedding geometries despite identical supervision. The SNN embeddings exhibit broader dispersion, whereas the MLP embeddings form relatively compact structures.

These findings suggest that temporal spiking dynamics influence representation topology even when retrieval performance remains comparable. The results highlight the importance of representation-level analysis in evaluating neuromorphic learning systems.

VI. Discussion

This study examined spiking and non-spiking neural architectures under a controlled multimodal contrastive learning framework. By maintaining identical architectures and optimization settings, temporal spiking dynamics were isolated as the primary experimental variable.

A. Stable Embeddings but Weak Retrieval

Both the Spiking Neural Network (SNN) and Multilayer Perceptron (MLP) learned stable and non-collapsed embedding spaces, yet retrieval performance remained close to chance level. Embedding norms and cosine similarity distributions confirmed that both models formed structured representations without numerical instability.

These results suggest that the primary limitation lies not in representational capacity, but in the mismatch between static supervision and multimodal retrieval requirements. Although contrastive learning produced distributed embeddings, the supervision signal was insufficient to establish strong semantic alignment.

B. Representation Divergence Between SNN and MLP

Despite identical training objectives and optimization settings, the SNN and MLP produced substantially different embedding geometries. Cross-model cosine similarity remained near zero, indicating limited representational overlap.

This divergence highlights the influence of architectural inductive bias on representation learning. Temporal membrane integration and spike accumulation in the SNN alter how semantic information is organized compared with static feed-forward computation in the MLP.

Importantly, these geometric differences did not improve retrieval accuracy, showing that similar task performance can conceal major representational differences.

C. Temporal Dynamics Under Static Supervision

The absence of retrieval improvement for the SNN is largely explained by the static nature of the dataset and supervision objective. Although spike encoding introduces temporal activity, the contrastive objective itself contains no temporal dependency requiring temporal reasoning.

Consequently, temporal spike dynamics are compressed into static embeddings before similarity computation. This limits the functional benefit of temporal computation under static multimodal supervision.

The findings therefore suggest a mismatch between temporally dynamic architectures and temporally unstructured learning objectives.

D. Implications for Neuromorphic Learning

The results indicate that the effectiveness of spiking architectures depends strongly on task structure and supervision design. Static image–text retrieval tasks may not fully exploit the computational advantages of SNNs.

Tasks involving sequential perception, event-driven sensing, or temporally evolving multimodal contexts may provide more suitable environments for spike-based representation learning.

More broadly, the study demonstrates that representation-level analysis is essential when evaluating neuromorphic systems. Architectures with similar retrieval performance can still

organize information using fundamentally different geometric structures.

VII. Limitations

This study was designed as a controlled representational analysis rather than a performance-oriented multimodal retrieval system. Although the controlled setup improves interpretability, several limitations affect the generalizability of the findings.

A. Static Supervision Constraints

The experiments were conducted on a static image–text retrieval benchmark without temporal dependencies or sequential structure. Consequently, the supervision objective does not explicitly require temporal reasoning, limiting the ability of the Spiking Neural Network (SNN) to exploit temporal dynamics effectively.

B. Limited Dataset Scale

The benchmark consists of only 1,000 multimodal pairs, which is relatively small compared with modern contrastive learning datasets. Limited semantic diversity may reduce the effectiveness of representation learning and multimodal alignment.

C. Restricted Architectural Scope

The study evaluates a single Leaky Integrate-and-Fire (LIF) spiking architecture against a Multilayer Perceptron (MLP) baseline. Alternative architectures such as recurrent spiking networks, transformers, or hybrid neuromorphic models may exhibit different representational behavior.

D. Rate-Based Spike Encoding

Temporal activity was generated using rate-based Poisson encoding applied to static features. Although effective for introducing spike-based computation, this method does not preserve precise spike timing or biologically realistic temporal dynamics.

E. Static Learning Objective

The contrastive objective operates on static embeddings and does not optimize temporal consistency or sequence prediction. This mismatch between temporal computation and static supervision may limit the functional advantages of spiking architectures.

F. Evaluation Limitations

The analysis primarily relies on Recall@K, MRR, cosine similarity statistics, PCA, and t-SNE visualization. While useful for representational analysis, these metrics provide only partial insight into embedding geometry and downstream transfer behavior.

G. Limited Generalizability

The findings are specific to a controlled experimental setting involving static multimodal supervision and a small benchmark dataset. Different results may emerge under larger-scale training, temporally structured datasets, or neuromorphic hardware implementations.

Nevertheless, the study provides a controlled empirical comparison of spiking and non-spiking representations under identical multimodal supervision conditions.

VIII. Conclusion

This study presented a controlled comparison between Spiking Neural Networks (SNNs) and Multilayer Perceptrons (MLPs) for multimodal image–text retrieval under static contrastive supervision. Both architectures were trained using identical embedding dimensions, optimization settings, and retrieval protocols to isolate the effect of temporal spiking dynamics.

Experimental results showed that both models learned stable and non-collapsed embedding spaces. However, retrieval performance remained close to random-chance levels, indicating limited multimodal alignment under the evaluated static benchmark. Despite similar task-level performance, representational analysis revealed substantial geometric divergence between SNN and MLP embeddings. Cross-model cosine similarity remained near zero, suggesting that both architectures organize multimodal information using fundamentally different representational structures.

The findings indicate that temporal spiking dynamics strongly influence embedding geometry, even when they do not improve retrieval effectiveness. The study also demonstrates that retrieval metrics alone are insufficient for evaluating biologically inspired neural systems. Embedding-level analyses, including cosine similarity statistics and low-dimensional visualization, provide important complementary insight into representation behavior.

More broadly, the results suggest a mismatch between temporal spiking computation and static multimodal supervision. Since the dataset lacks explicit temporal structure, the advantages of spike-based temporal processing may remain underutilized. Future work may investigate temporally structured multimodal tasks, event-driven learning environments, recurrent spiking architectures, and alternative spike encoding strategies to better understand the role of temporal dynamics in neuromorphic representation learning.

References

[1] W. Maass, “Networks of spiking neurons: The third generation of neural network models,” *Neural Networks*, vol. 10, no. 9, pp. 1659–1671, 1997.
[2] E. O. Neftci, H. Mostafa, and F. Zenke, “Surrogate gradient learning in spiking neural networks,” *IEEE Signal Processing Magazine*, vol. 36, no. 6, pp. 61–63, 2019.
[3] P. U. Diehl and M. Cook, “Unsupervised learning of digit recognition using spike-timing-dependent plasticity,” *Frontiers in Computational Neuroscience*, vol. 9, 2015.
[4] S. Roy, A. Banerjee, and A. Basu, “Toward spike-based machine intelligence with neuromorphic computing,” *Nature*, vol. 575, pp. 607–617, 2019.

[5] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2020, pp. 1597–1607.
[6] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2021, pp. 8748–8763.
[7] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.
[8] J. Deng *et al.*, "ImageNet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 248–255.
[9] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.
[10] J. Johnson, M. Douze, and H. Je´gou, "Billion-scale similarity search with FAISS," *IEEE Transactions on Big Data*, 2019.
[11] B. Rueckauer *et al.*, "Conversion of continuous-valued deep networks to efficient event-driven networks," *Frontiers in Neuroscience*, vol. 11, 2017.
[12] F. Zenke and T. P. Vogels, "The remarkable robustness of surrogate gradient learning for instilling complex function in spiking neural networks," *Neural Computation*, vol. 33, no. 4, pp. 899–925, 2021.
[13] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
[14] A. Graves, *Supervised Sequence Labelling with Recurrent Neural Networks*. Berlin, Germany: Springer, 2012.
[15] D. Hubel and T. Wiesel, "Receptive fields of single neurons in the cat's striate cortex," *Journal of Physiology*, vol. 148, no. 3, pp. 574–591, 1959.
[16] H. Markram, "The blue brain project," *Nature Reviews Neuroscience*, vol. 7, no. 2, pp. 153–160, 2006.
[17] M. Davies *et al.*, "Loihi: A neuromorphic manycore processor with on-chip learning," *IEEE Micro*, vol. 38, no. 1, pp. 82–99, 2018.
[18] C. Eliasmith *et al.*, "A large-scale model of the functioning brain," *Science*, vol. 338, no. 6111, pp. 1202–1205, 2012.
[19] Y. Bengio, "Learning deep architectures for AI," *Foundations and Trends in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.
[20] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
[21] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, 2015.
[22] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proceedings of the Neural Information Processing Systems Conference (NeurIPS)*, 2012, pp. 1097–1105.
[23] A. Dosovitskiy *et al.*, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.
[24] J. Devlin *et al.*, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.