SpikingVTG: A Spiking Detection Transformer for Video Temporal Grounding

Malyaban Bal[1,2], Brian Matejek[1], Susmit Jha[1], Adam Cobb[1]
SRI International[1], Penn State University[2]

Introduction

The rapid growth of social media and portable devices has led to an explosion of video content, creating a pressing need for efficient methods to **retrieve** and **analyze** video data. As a result, **Video Temporal Grounding** (VTG) [1] has become a key area of research. Although foundation models such as **LLMs** and **VLMs** have boosted VTG performance, their heavy computational and energy demands hinder practical real-world deployment. In this paper we propose a **first-of-its-kind spiking** framework for VTG. We introduce a **Saliency Feedback Gating** mechanism that leverages temporal dynamics in SNNs to enhance performance and reduce computational overhead. Additionally, we propose Cos-L2 Representation Matching (CLRM) for efficient training and multiple optimization strategies for developing on-chip deployable model variants suitable for resource-constrained environments.

Primary Contributions

- **SpikingVTG Model and Training Framework:** We propose a transformer-based, multimodal spiking video language model for moment retrieval and highlight detection in VTG tasks. We leverage the layer-wise convergence dynamics in our model to train our model using implicit differentiation at equilibrium, bypassing memory intensive BPTT. The result is the first spiking model demonstrating competitive performance on VTG.
- **Saliency Feedback Gating Mechanism**: We introduce a saliency feedback gating mechanism for input video, that leverages the ASR of the output of the spiking transformer. This temporal feedback enhances task-specific performance while minimizing neural activity, reducing overall computational overhead.
- **Efficient Training and Inference Optimizations:** We propose **Cos-L2 Representation Matching (CLRM),** a lightweight knowledge-transfer objective that distills non-spiking teachers into spiking models. In addition, we develop a **normalization-free 1-bit SpikingVTG** variant tailored for **low-power, edge-friendly inference**.

Layer-wise Dynamics

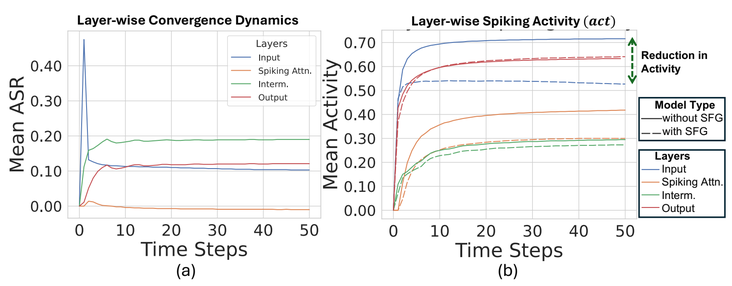

Figure 1. (a) Graph shows convergence dynamics of layer-wise mean ASR against operating time steps for a randomly selected spiking transformer encoder layer (Fig. 1). (b) Graph shows layer-wise mean spiking activity against operating time steps in x-axis. The model with SFG shows markedly reduced activity in both the input layer and the spiking attention layer, underscoring its role in minimizing neuronal activity.

- Underlying LIF models uses ternary spikes, $s \in \{-1, 0, 1\}$.
- ASR convergence Dynamics at Equilibrium: $a_i^* = \sigma(\frac{1}{V_{th_i}}(\hat{f}(a_{i-1}^*) + b_i))$
- During training instead of BPTT, only ASR values at equilibrium are used to train the model using IDE [2].

KD and On-Chip Friendly Optimizations

- We leverage the layer-wise convergence dynamics to perform Knowledge transfer leveraging Cos-L2 Representation Matching (CLRM) from a ”teacher” ANN-based **UniVTG** model to our “student” **SpikingVTG** model.

$$\mathcal{L}_{\text{rep}} = \frac{1}{B \times L} \sum_{i=1}^{N} \sum_{j=1}^{B} \sum_{k=1}^{L} \left[\lambda_{\cos} \cdot \left(1 - \cos\left(\theta_{i,j,k}^{\text{rep}}\right)\right)^2 + \lambda_{\ell_2} \cdot \left\| \mathbf{s}_i^{(j,k)} - \mathbf{t}_i^{(j,k)} \right\|_2^2 \right] \quad \begin{aligned} \mathbf{t}_i^{(j,k)} &= T_{r_i}^{(j,k)} \in \mathbb{R}^{d_t} \\ \mathbf{s}_i^{(j,k)} &= a_{r_i}^{*(j,k)} W_d \in \mathbb{R}^{d_t} \end{aligned}$$

- To make SpikingVTG **neuromorphic**-friendly, we eliminate all **layer normalization** and substitute **softmax** in attention and SFG with a ReLU followed by a scaling term (L^{-1}), enabling efficient on-chip implementation.
- To **minimize** the **memory** footprint of our model for deployment on resource-constrained devices, we employ **extreme 1-bit quantization**.

SpikingVTG Architecture

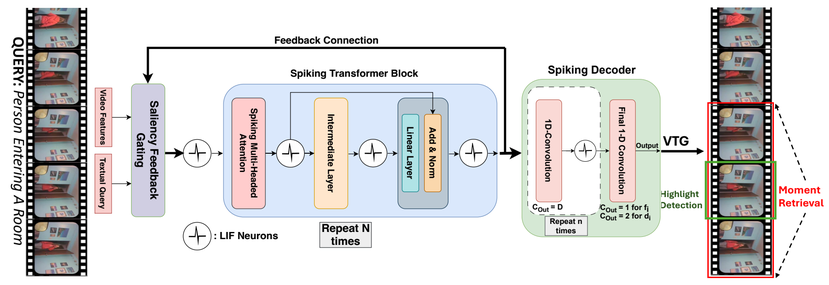

Figure 2. High-level overview of the SpikingVTG architecture. The spiking Vision- Language Model (VLM) takes video and textual features as inputs, employing a spiking transformer core that utilizes Saliency Feedback Gating through temporal feedback connections.

SFG Mechanism

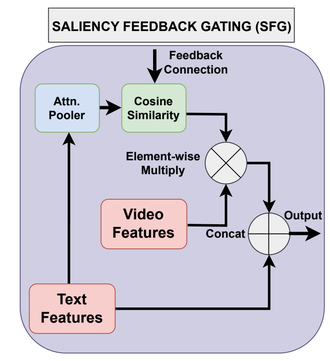

Figure 3. Overview of the internal operations of the SFG mechanism. The ASR of the output of the spiking transformer core at each time step is leveraged as the feedback signal.

Multi-Stage Training Pipeline

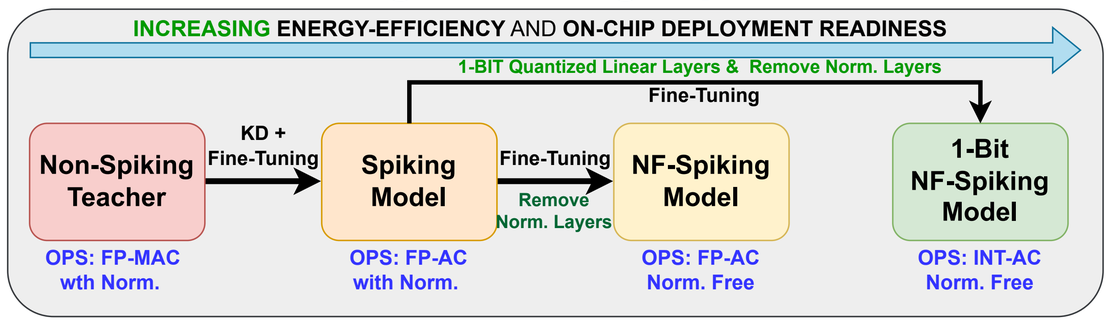

Figure 4. High-level overview of the multi-stage training framework for our proposed SpikingVTG models, enabling the development of lightweight and computationally efficient spiking models. Below each model we have noted the primary operations involved in that architecture.

Experimental Results

Method	SNN	QVHighlights-MR				Charades-STA			
		@0.5	@0.7	mAP@0.5	mAP@0.75	@0.3	@0.5	@0.7	mIoU
M-DETR [1]	No	52.89	33.02	54.82	29.40	65.83	52.07	30.59	45.54
UMT [5]	No	56.23	41.18	53.83	37.01	-	49.35	26.16	-
UniVTG [6]	No	58.86	40.86	57.60	35.59	70.81	58.01	35.65	50.10
UniVTG+PT [6]	No	65.43	50.06	64.06	45.02	72.63	60.19	38.55	52.17
UVCOM [34]	No	63.55	47.47	63.37	42.67	-	56.69	34.76	-
SpikeMba [19]	No	64.13	49.42	-	43.67	71.24	59.65	36.12	51.74
BAM-DETR [35]	No	62.71	48.64	64.57	46.33	72.93	59.95	39.38	52.33
TR-DETR [36]	No	64.66	48.96	63.98	43.73	-	57.61	33.52	-
CG-DETR [37]	No	65.43	48.38	64.51	42.77	70.40	58.40	36.30	50.10
LLMEPET [38]	No	66.73	49.94	65.76	43.91	70.91	-	36.49	50.25
SpikingVTG	**Yes**	**65.29**	**48.18**	**64.31**	**42.25**	**71.20**	**58.73**	**37.16**	**50.62**

Table 1: Performance comparison of our SpikingVTG model with SFG against non-spiking VTG solutions on the test set of the QVHighlights and Charades-STA for moment retrieval task.

Method	SNN	QVHighlights-HD	
		mAP	HIT@1
UMT [5]	No	38.18	59.99
UniVTG [6]	No	38.20	60.96
UniVTG+PT [6]	No	40.54	66.28
QD-DETR [39]	No	38.90	62.40
SpikeMba [5]	No	-	-
CG-DETR [37]	No	40.30	66.20
UVCOM [34]	No	39.98	65.58
LLMEPET [38]	No	38.18	59.99
SpikingVTG	**Yes**	**40.46**	**65.82**

Table 2: Performance comparison of our SpikingVTG model on the evaluation set of the QVHighlights for highlight detection task.

Method	QVHighlights-MR		QVHighlights-HD		Operations	Local	Activity	Energy
	@0.5	@0.7	mAP	HIT@1				
Pre-trained UniVTG (sota)	67.35	52.65	41.34	68.77	FP-MAC	×	1.0	23.92mJ
SpikingVTG **without SFG**	64.94	47.21	40.49	67.37	FP-ACC	×	0.41	15.2mJ
SpikingVTG **with SFG**	**67.58**	**50.82**	**40.81**	**68.64**	FP-ACC	×	0.34	13.8mJ
(NF)-SpikingVTG w/ SFG	66.59	48.31	40.61	67.73	FP-ACC	✓	0.25	10.1mJ
1-bit (NF)-SpikingVTG w/ SFG	65.31	47.48	40.35	67.30	**INT-ACC**	✓	**0.19**	**1.3mJ**
1-bit (NF)-SpikingVTG w/ ReLU	65.91	47.04	40.16	67.07	**INT-ACC**	✓	**0.19**	**1.3mJ**

Table 3: Performance comparison of the different SpikingVTG variants as highlighted in Fig. 4 on the evaluation set of QVHighlights dataset.

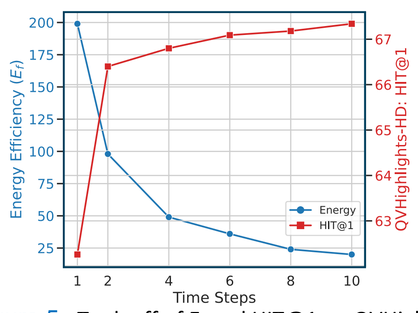

Figure 5.: Tradeoff of E_f and HIT@1 on QVHighlights-HD for varying time steps.

- First **spiking framework** for **VTG**, enabling moment retrieval and highlight detection.
- **SFG** mechanism results in **improved performance** and **reduced computational overhead**.
- **1-bit NF SpikingVTG** achieves minimal performance degradation despite removing non-local normalization and employing extreme 1-bit weight quantization, making it the most hardware-friendly implementation for **edge devices**.

Conclusions

- The SFG mechanism proposed in this work capitalizes on the temporal dynamics of SNNs to adaptively process input video. This approach enhances performance while reducing computational overhead.
- We train SpikingVTG using IDE with equilibrium convergence dynamics, employ CLRM for efficient knowledge transfer, and develop 1-bit (NF)-SpikingVTG variant tailored for resource-constrained devices.
- **Future Work**: Deploy SpikingVTG on neuromorphic hardware (e.g., Intel Loihi 2) and integrate with event-based video inputs from spiking cameras.

References

[1] Jie Lei, Tamara L Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. In NeurIPS, pages 11846–11858, 2021.

[2] Mingqing Xiao, Qingyan Meng, Zongpeng Zhang, Yisen Wang, and Zhouchen Lin. Training feedback spiking neural networks by implicit differentiation on the equilibrium state. Advances in Neural Information Processing Systems, 34:14516–14528, 2021.

Acknowledgement

This work was supported in part by the United States Air Force and Defense Advanced Research Projects Agency (DARPA) under Contract No. FA8750-23-C-0519 and the U.S. Army Research Laboratory Cooperative Research Agreement W911NF-17-2-0196. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the Department of Defense or the United States Government.