Enhancing Partially Relevant Video Retrieval with Hyperbolic Learning

Supplementary Material

A. Appendix

A.1. Derivation of L_{pop}

In this section, we formally derive the components of the partial order preservation loss employed in our approach.

Half-Aperture We begin with the definition of the half-aperture for the Poincar'e ball, as introduced by Ganea et al. [3]. Given a point $\boldsymbol{x}_{\mathcal{P}}$ on the Poincar'e ball, the cone half-aperture is formulated as:

$$\mathrm{HA}_{\mathcal{P}}(\boldsymbol{x}_{\mathcal{P}}) = \sin^{-1}\left(c\frac{1-\|\boldsymbol{x}_{\mathcal{P}}\|^2}{\|\boldsymbol{x}_{\mathcal{P}}\|}\right). \tag{22}$$

Since the Poincar'e ball model and the Lorentz hyperboloid model are isometric, any point $\boldsymbol{x}_{\mathcal{P}}$ in the Poincar'e ball can be mapped to a corresponding point $\boldsymbol{x}_{\mathcal{L}}$ in the hyperboloid model via the following differentiable transformation:

$$\boldsymbol{x}_{\mathcal{L}} = \frac{2\boldsymbol{x}_{\mathcal{P}}}{1-\|\boldsymbol{x}_{\mathcal{P}}\|^2}. \tag{23}$$

To ensure model invariance, the half-aperture should remain unchanged across hyperbolic representations, i.e., $\mathrm{HA}_{\mathcal{L}}(\boldsymbol{x}_{\mathcal{L}}) = \mathrm{HA}_{\mathcal{P}}(\boldsymbol{x}_{\mathcal{P}})$. Substituting Eq. (23) into Eq. (22), we derive:

$$\mathrm{HA}_{\mathcal{L}}(\boldsymbol{x}_{\mathcal{L}}) = \sin^{-1}\left(\frac{2c}{\|\boldsymbol{x}_{\mathcal{L}}\|}\right). \tag{24}$$

Exterior Angle Consider three points: the origin $\boldsymbol{o}$, the video embedding $\boldsymbol{v}$, and the text embedding $\boldsymbol{t}$. These points form a hyperbolic triangle whose sides are defined by the geodesic distances $x = d^2_{\mathcal{L}}(\boldsymbol{o},\boldsymbol{t})$, $y = d^2_{\mathcal{L}}(\boldsymbol{o},\boldsymbol{v})$, and $z = d^2_{\mathcal{L}}(v,\boldsymbol{t})$. The hyperbolic law of cosines provides a means to compute the angles of this triangle. The exterior angle is given by:

$$\begin{aligned}\mathrm{EA}(\boldsymbol{v},\boldsymbol{t}) &= \pi - \angle \boldsymbol{ovt} \\ &= \pi - \cos^{-1}\left[\frac{\cosh(z)\cosh(y)-\cosh(x)}{\sinh(z)\sinh(y)}\right].\end{aligned} \tag{25}$$

We define $g(s) = \cosh(s)$ and employ the hyperbolic identity $\sinh(s) = \sqrt{\cosh^2(s)-1}$:

$$\mathrm{EA}(\boldsymbol{v},\boldsymbol{t}) = \cos^{-1}\left[\frac{g(x)-g(z)g(y)}{\sqrt{g(z)^2-1}\sqrt{g(y)^2-1}}\right]. \tag{26}$$

We now compute $g(x)$, $g(y)$, and $g(z)$. Given that $g(z) = \cosh(d^2_{\mathcal{L}}(\boldsymbol{v},\boldsymbol{t}))$ and utilizing the definition $d^2_{\mathcal{L}}(\boldsymbol{v},\boldsymbol{t}) = \cosh^{-1}(-\langle \boldsymbol{v},\boldsymbol{t}\rangle_{\mathcal{L}})$, we obtain:

$$\begin{aligned} g(z) &= \cosh\left(d^2_{\mathcal{L}}(\boldsymbol{v},\boldsymbol{t})\right) \\ &= \cosh\left(\cosh^{-1}(-\langle \boldsymbol{v},\boldsymbol{t}\rangle_{\mathcal{L}})\right) \\ &= -\langle \boldsymbol{v},\boldsymbol{t}\rangle_{\mathcal{L}}. \end{aligned} \tag{27}$$

Similarly, we derive $g(x) = -\langle \boldsymbol{o},\boldsymbol{t}\rangle_{\mathcal{L}}$ and $g(y) = -\langle \boldsymbol{o},\boldsymbol{v}\rangle_{\mathcal{L}}$. The Lorentzian inner product with the origin $\boldsymbol{o}$ simplifies as follows:

$$\langle \boldsymbol{o},\boldsymbol{v}\rangle_{\mathcal{L}} = -v_0, \quad \text{and} \quad \langle \boldsymbol{o},\boldsymbol{t}\rangle_{\mathcal{L}} = -t_0. \tag{28}$$

Thus, we obtain $g(x) = t_0$ and $g(y) = v_0$. Substituting these values into Eq. (26), we derive the refined expression:

$$\mathrm{EA}(\boldsymbol{v},\boldsymbol{t}) = \cos^{-1}\left(\frac{t_0 + v_0\langle v,t\rangle\mathcal{L}}{\sqrt{v_0^2-1}\sqrt{(\langle v,t\rangle\mathcal{L})^2-1}}\right).$$

Finally, utilizing the relation between x_0 and vs, we simplify the denominator to obtain the final expression for the exterior angle:

$$\mathrm{EA}(\boldsymbol{v},\boldsymbol{t}) = \cos^{-1}\left(\frac{t_0 + v_0\langle v,t\rangle\mathcal{L}}{\|\boldsymbol{v}_s\|\sqrt{(\langle v,t\rangle\mathcal{L})^2-1}}\right).$$

A.2. Training Objectives

Following existing works [2, 8], we adopt triplet loss [1, 5] L^{trip} and InfoNCE loss [4, 6, 9, 10] L^{nce}, query diverse loss [7, 8] L_{div}. A text-video pair is considered positive if the video contains a moment relevant to the text; otherwise, it is regarded as negative. Given a positive text-video pair $(\mathcal{T},\mathcal{V})$, the triplet ranking loss over the mini-batch $\mathcal{B}$ is formulated as:

$$\begin{aligned} L^{trip} = \frac{1}{N}\sum_{(\mathcal{T},\mathcal{V})\in\mathcal{B}} \{&\max(0, m + S(\mathcal{T}^-,\mathcal{V}) - S(\mathcal{T},\mathcal{V})) \\ &+\max(0, m + S(\mathcal{T},\mathcal{V}^-) - S(\mathcal{T},\mathcal{V}))\}, \end{aligned} \tag{29}$$

where m is a margin constant. $\mathcal{T}^-$ and $\mathcal{V}^-$ indicate a negative text for $\mathcal{V}$ and a negative video for $\mathcal{T}$, respectively. The similarity score $S(,)$ is obtained by Equation (9) .

The infoNCE loss is computed as:

$$\begin{aligned} L^{nce} = -\frac{1}{N}\sum_{(\mathcal{T},\mathcal{V})\in\mathcal{B}} \{&\log(\frac{S(\mathcal{T},\mathcal{V})}{S(\mathcal{T},\mathcal{V}) + \sum_{\mathcal{T}_i^-\in\mathcal{N}_{\mathcal{T}}} S(\mathcal{T}_i^-,\mathcal{V})}) \\ &+\log(\frac{S(\mathcal{T},\mathcal{V})}{S(\mathcal{T},\mathcal{V}) + \sum_{\mathcal{V}_i^-\in\mathcal{N}_{\mathcal{V}}} S(\mathcal{T},\mathcal{V}_i^-)})\}, \end{aligned} \tag{30}$$

where $\mathcal{N}_{\mathcal{T}}$ and $\mathcal{N}_{\mathcal{V}}$ represent the negative texts and videos of $\mathcal{V}$ and $\mathcal{T}$ within the mini-batch $\mathcal{B}$, respectively.

Finally , L_{sim} is defined as:

$$L_{sim} = L^{trip}_{clip} + L^{trip}_{frame} + \lambda_c L^{nce}_{clip} + \lambda_f L^{nce}_{frame}, \tag{31}$$

Name	Configuration
CPU	Intel® Xeon® Platinum 8269CY CPU @ 2.50GHz (26 cores)
GPU	A single NVIDIA GeForce GTX 3080 Ti (12GB)
RAM	64GB
OS	Ubuntu 20.04 LTS
CUDA Version	11.7
GPU Driver Version	535.183.01
Language	Python 3.11.8
Dependencies	torch 2.0.1 torchvision 0.15.2 numpy 1.26.4

Table 3. Computing infrastructure for our experiments.

Params	ActivityNet Captions	TVR	Charades-STA
learning rate	2.5e-4	3e-4	2e-4
α_f	0.3	0.3	0.3
α_c	0.7	0.7	0.7
α	32	32	32
δ	0.2	0.15	0.2
m	0.2	0.1	0.2
τ	6e-1	9e-2	6e-1
λ_c	2e-2	5e-2	2e-2
λ_f	4e-2	4e-2	4e-2
λ_1	3e-3	8e-5	3e-3
λ_2	1e-3	1e-3	1e-3

Table 4. Hyper-parameter settings.

where $frame$ and $clip$ mark the objectives for the gaze frame-level branch and the glance clip-level branch, respectively. λ_c and λ_f are hyper-parameters to balance the contributions of InfoNCE objectives.

Given a collection of text queries $\mathcal{T}$ in the mini-batch $\mathcal{B}$, the query diverse loss is defined as:

$$L_{div} = \frac{1}{N} \sum_{q_i,q_j \in \mathcal{T}} \mathbb{1}_{q_i,q_j} log(1 + e^{\alpha(cos(q_i,q_j)+\delta)}) \quad (32)$$

where $\delta > 0$ denotes the margin, $\alpha > 0$ is a scaling factor, and $\mathbb{1}_{q_i,q_j} \in \{0, 1\}$ represents an indicator function, $\mathbb{1}_{q_i,q_j} = 1$ when q_i and q_j correspond to the same video.

B. Experiments

B.1. Details of Experimental Setup

Details of Training Configurations The computing infrastructure is in Table 3. All random seeds are set to 0.

Hyper-parameter Notably, we directly inherent most hyper-parameter settings from GMMFormer. In detail, we use $M_c = 32$ for downsampling and set the maximum frame number $M_f = 128$. If the number of frames exceeds M_f, we uniformly downsample it to M_f. For sentences, we set the maximum length of query words to $N_q = 64$ for ActivityNet Captions and $N_q = 30$ for TVR and Charades-STA. Any words beyond the maximum length will be discarded. The Lorentz latent dimension $n = 127$. You can find other detailed hyper-parameter settings in Tab. 4.

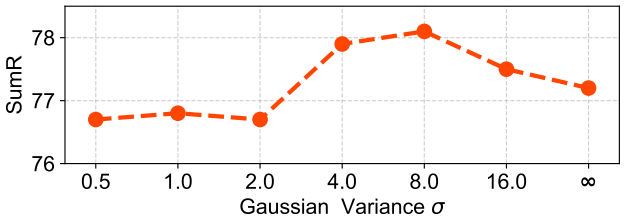

Figure 6. The impact of the Gaussian variance σ on Charades-STA.

B.2. Additional Results on Model Analyses

Impact of the Gaussian Variance σ We investigate the impact of the Gaussian variance σ on experimental results by employing a uniform σ across all Gaussian attention blocks. As illustrated in Fig. 6, larger σ generally leads to superior performance due to its broader receptive field, which enables better modeling of temporal dependencies within videos. However, excessively large σ results in overly dispersed attention, weakening the enhancement of semantic information from adjacent frames or clips, thereby leading to suboptimal performance. In contrast, HLFormer employs multiple σ values to achieve multi-scale flexible of video semantics, not only attaining improved performance but also mitigating the need for extensive hyper-parameter tuning.

References

[1] Jianfeng Dong, Xirong Li, Chaoxi Xu, Xun Yang, Gang Yang, Xun Wang, and Meng Wang. Dual encoding for video retrieval by text. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4065–4080, 2021. 1

[2] Jianfeng Dong, Xianke Chen, Minsong Zhang, Xun Yang, Shujie Chen, Xirong Li, and Xun Wang. Partially relevant video retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 246–257, 2022. 1

[3] Octavian Ganea, Gary Becigneul, and Thomas Hofmann. Hyperbolic entailment cones for learning hierarchical embeddings. In *Proceedings of the 35th International Conference on Machine Learning*, pages 1646–1655. PMLR, 2018. 1

[4] Guanghao Meng, Sunan He, Jinpeng Wang, Tao Dai, Letian Zhang, Jieming Zhu, Qing Li, Gang Wang, Rui Zhang, and Yong Jiang. Evdclip: Improving vision-language retrieval with entity visual descriptions from large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6126–6134, 2025. 1

[5] Haomiao Tang, Jinpeng Wang, Yuang Peng, GuangHao Meng, Ruisheng Luo, Bin Chen, Long Chen, Yaowei Wang, and Shu-Tao Xia. Modeling uncertainty in composed image retrieval via probabilistic embeddings. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1210–1222, 2025. 1

[6] Yuanmin Tang, Jing Yu, Keke Gai, Jiamin Zhuang, Gang Xiong, Gaopeng Gou, and Qi Wu. Missing target-relevant information prediction with world model for accurate zero-shot composed image retrieval. In *Proceedings of the Computer*

Vision and Pattern Recognition Conference, pages 24785–24795, 2025. 1

[7] Yuting Wang, Jinpeng Wang, Bin Chen, Tao Dai, Ruisheng Luo, and Shu-Tao Xia. Gmmformer v2: An uncertainty-aware framework for partially relevant video retrieval, 2024. 1

[8] Yuting Wang, Jinpeng Wang, Bin Chen, Ziyun Zeng, and Shu-Tao Xia. Gmmformer: Gaussian-mixture-model based transformer for efficient partially relevant video retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024. 1

[9] Hao Zhang, Aixin Sun, Wei Jing, Guoshun Nan, Liangli Zhen, Joey Tianyi Zhou, and Rick Siow Mong Goh. Video corpus moment retrieval with contrastive learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 685–695, 2021. 1

[10] Minyi Zhao, Jinpeng Wang, Dongliang Liao, Yiru Wang, Huanzhong Duan, and Shuigeng Zhou. Keyword-based diverse image retrieval by semantics-aware contrastive learning and transformer. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1262–1272, 2023. 1