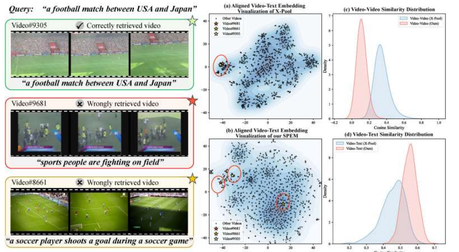
Similarity-aware Probabilistic Embeddings Modeling for Video-Text Retrieval

Yuliang Huang[1], Pengxu Wei[1,2*], Zhicheng Dong[3], Lin Liang[1,2]
[1]Sun Yat-sen University, [2]Peng Cheng Laboratory, [3]Tibet University

WACV 2026 TUCSON, AZ 3/6 – 3/10

Abstract

Motivation:

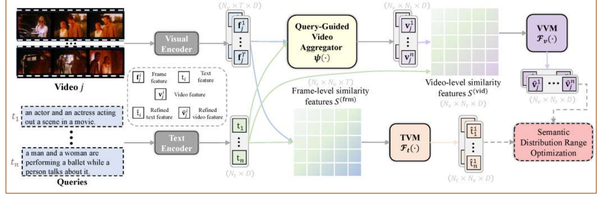


- Modern video-text retrieval methods embed videos and texts as deterministic vectors. In practice, visually dominant patterns such as recurring backgrounds or common objects can cause semantically different videos to collapse into nearby regions of the embedding space.
- This semantic over-clustering suppresses fine-grained distinctions and blurs the boundary between relevant samples and hard negatives, ultimately limiting retrieval precision—especially for short or ambiguous queries.
- Core Idea.
  - Capture semantic uncertainty,
  - Reduce embedding collapse
  - Improve discriminative alignment

Methods and Materials

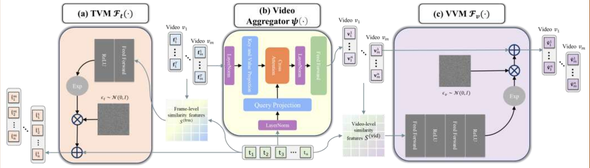
SPEM Overview:

- SPEM enhances cross-modal alignment by modeling video and text embeddings as semantic distributions rather than deterministic vectors.

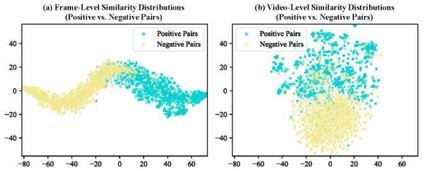


Methods and Materials

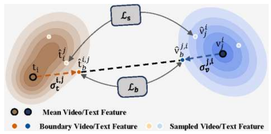
Modules design:



- Cross-Modal Attentive Fusion
  - A query-guided transformer aggregates frame features to highlight text-relevant content while suppressing visually dominant but irrelevant patterns.
- Uncertainty-aware Distribution Projection
  - Embeddings are projected into Gaussian distributions whose means remain anchored to deterministic features, preventing semantic drift while enabling controlled variability.
- Similarity-Guided Variance Adaptation
  - Distribution variance is dynamically predicted from frame-level and video-level similarity signals, allowing the embedding space to flexibly expand or contract based on alignment quality.



- Semantic Distribution Contrastive Loss
  - We optimize sampled and boundary embeddings using a symmetric contrastive objective, improving separation between positives and hard negatives.



Results

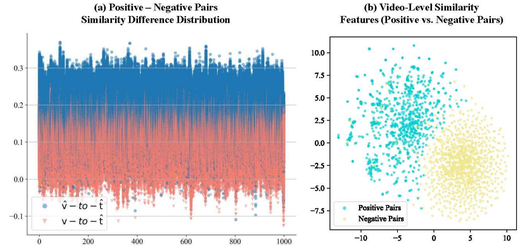
State-of-the-Art Performance:

- SPEM consistently improves retrieval accuracy across multiple datasets while maintaining efficiency comparable to existing CLIP-based approaches.

Method	Video → Text			Text → Video		
	R@1 ↑	R@5 ↑	MnR ↓	R@1 ↑	R@5 ↑	MnR ↓
X-Pool	44.4	73.3	9.0	46.9	72.8	14.3
ProST	46.3	74.2	8.7	48.2	74.6	12.4
UATVR	46.9	73.8	8.8	47.5	73.9	12.3
Cap4Video	47.1	73.7	8.7	49.3	74.3	12.0
DiffusionRet	47.7	73.8	8.8	49.0	75.2	12.1
KDProR	47.2	74.6	8.7	48.7	74.4	12.0
MUSE	49.7	77.8	7.4	50.9	76.7	10.9
ours	**54.1**	**83.8**	**5.6**	**51.3**	**79.9**	**7.8**

Method	DiDeMo			MSVD			VATEX		
	R@1	R@5	MnR	R@1	R@5	MnR	R@1	R@5	MnR
CLIP4Clip	42.8	68.5	18.9	45.2	75.5	10.3	55.9	89.2	3.9
X-Pool	44.6	73.2	15.4	47.2	77.4	9.3	60.0	90.0	3.8
DRL	47.9	73.8	-	48.3	79.1	-	63.5	91.7	-
UATVR	43.1	71.8	15.1	46.0	76.3	10.4	61.3	91.0	3.3
ours	**50.9**	**79.3**	**10.7**	**50.5**	**80.7**	**7.1**	**64.8**	**93.4**	**2.0**

- Probabilistic embeddings enlarge the similarity gap between positive and negative pairs, reducing semantic overlap and improving alignment robustness.



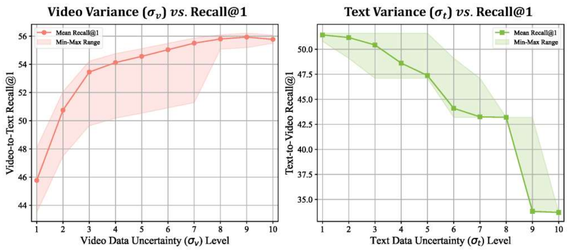
Discussion

Efficiency:

- SPEM introduces only lightweight sampling and small neural modulators, resulting in nearly identical GFLOPs to strong baselines while delivering significant gains in Recall@1.

Method	GFLOPs	MSRVTT		DiDeMo		MSVD	
		L(s)	R@1	L(s)	R@1	L(s)	R@1
CLIP4Clip	54.427	0.105	44.5	0.029	42.8	0.025	45.2
X-Pool	53.236	0.107	46.9	0.032	47.9	0.028	47.2
UATVR	55.063	0.184	47.5	0.061	43.1	0.048	46.0
Ours	**53.237**	**0.126**	**51.3**	**0.053**	**50.9**	**0.040**	**50.5**

- Modeling embeddings as adaptive distributions provides a flexible mechanism for capturing semantic diversity without sacrificing alignment stability. Moderate video uncertainty improves retrieval by mitigating over-clustering, while controlled text variance preserves semantic precision. These findings support modality-specific uncertainty design for cross-modal representation learning.



Conclusions

Summary/Conclusion

- We introduced SPEM, a similarity-aware probabilistic framework that models video and text embeddings as adaptive semantic distributions. By dynamically modulating variance and enforcing boundary-aware alignment, SPEM mitigates semantic over-clustering and enhances cross-modal discriminability.
- Extensive evaluations on MSRVTT, DiDeMo, MSVD, and VATEX show that SPEM consistently outperforms strong CLIP-based baselines while incurring minimal computational overhead, offering an effective paradigm for uncertainty-aware cross-modal representation learning.

Contact

Yuliang Huang
Sun Yat-sen University
Email: huangyliang5@mail2.sysu.edu.cn
Website: https://www.sysu-hcp.net/home/

References

1. Luo, H., Ji, L., Zhong, M., Chen, Y., Lei, W., Duan, N., & Li, T. (2022). Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. Neurocomputing, 508, 293-304.
2. Gorti, S. K., Vouitsis, N., Ma, J., Golestan, K., Volkovs, M., Garg, A., & Yu, G. (2022). X-pool: Cross-modal language-video attention for text-video retrieval. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 5006-5015).
3. Li, P., Xie, C. W., Zhao, L., Xie, H., Ge, J., Zheng, Y., ... & Zhang, Y. (2023). Progressive spatio-temporal prototype matching for text-video retrieval. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 4100-4110).
4. Fang, B., Wu, W., Liu, C., Zhou, Y., Song, Y., Wang, W., ... & Wang, J. (2023). Uatvr: Uncertainty-adaptive text-video retrieval. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 13723-13733).
5. Wu, W., Luo, H., Fang, B., Wang, J., & Ouyang, W. (2023). Cap4video: What can auxiliary captions do for text-video retrieval? In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10704-10713).
6. Jin, P., Li, H., Cheng, Z., Li, K., Ji, X., Liu, C., ... & Chen, J. (2023). Diffusionret: Generative text-video retrieval with diffusion model. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 2470-2481).
7. Zhuang, X., Li, H., Cheng, X., Zhu, Z., Xie, Y., & Zou, Y. (2024, September). Kdpror: A knowledge-decoupling probabilistic framework for video-text retrieval. In European Conference on Computer Vision (pp. 313-331). Cham: Springer Nature Switzerland.
8. Wang, Q., Zhang, Y., Zheng, Y., Pan, P., & Hua, X. S. (2022). Disentangled representation learning for text-video retrieval. arXiv preprint arXiv:2203.07111.
9. Yang, H., Cao, M., Huang, J., Liu, R., Jin, P., Li, C., & Liang, X. (2025, April). Muse: Mamba is efficient multi-scale learner for text-video retrieval. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 39, No. 7, pp. 7238-7246).
10. Chun, S., Oh, S. J., De Rezende, R. S., Kalantidis, Y., & Larlus, D. (2021). Probabilistic embeddings for cross-modal retrieval. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 8415-8424).