Supplementary Material of "Text Is MASS: Modeling as Stochastic Embedding for Text-Video Retrieval"

Jiamian Wang[1], Guohao Sun[1], Pichao Wang[2]*, Dongfang Liu[1]†,
Sohail Dianat[1], Majid Rabbani[1], Raghuveer Rao[3], Zhiqiang Tao[1]†
[1]Rochester Institute of Technology, [2]Amazon Prime Video, [3]Army Research Laboratory

We provide more results, visualizations, and in-depth discussions of the proposed T-MASS as follows

- More quantitative performance of T-MASS (Section 1).
- Discussions about stochastic text embedding (Section 2).
- Text length analysis of T-MASS (Section 3)
- More discussions on KL divergence (Section 4).

1. Performance

Post Processing Operation. We further explore the power of propose T-MASS by evaluating upon post processing operation of DSL [4]. In Table 1, we provide results under the backbone of CLIP-ViT-B/32. In Table 2, we provide results under CLIP-ViT-B/16. Consistent performance boost can be achieved under different datasets. For example, DSL enables 4.1% improvement on R@1 for DiDeMo [1] with CLIP-ViT-B/32 and 8% improvement on R@1 for VATEX [13] with CLIP-ViT-B/16. Besides, we notice that the effect of DSL varies on different datasets and backbones, for which reason, we did not consider the post processing operations such as DSL and QB-Norm [3] when compare between different methods in the manuscript.

Video-to-text performance. In Table 4, we report both text-to-video and video-to-text retrieval performance of the proposed T-MASS with the backbone of CLIP-ViT-B/32. For some datasets, such as DiDeMo [1] and Charades [12], the video-to-text performance of the proposed method is close to text-to-video performance, *e.g.*, 1% gap on R@1. For others, such as MSRVTT [14] and VATEX [13], video-to-text performance of T-MASS can be even better than text-to-video performance, notably, 13.4% gap on R@1 for VATEX. Both of observations indicates that the proposed method not only allows a promising text-to-video retrieval, but also enables an accurate video-to-text retrieval. This is because T-MASS learns the text embedding as a resilient and flexible semantic range, facilitating the text-video representation alignment in the embedding space.

Larger Settings. In Table 5, we provide text-to-video performance of T-MASS with a larger batch size of 64 on DiDeMo and LSMDC. Better performance is achieved. Note that in the manuscript, we report the performance of the proposed T-MASS with batch size of 32 and frame number of 12 for all datasets. We directly report the performance of other methods under their own settings. For example, CLIP-ViP [15] uses batch size of 128 for different datasets and UATVR [5] use batch size of 64 during training. The comparison in the manuscript can be unfair to the proposed method. Despite the inconsistent setting, T-MASS achieves the state-of-the-art performance in most cases, validating the effectiveness of the proposed design.

2. Stochastic Text Embedding

Implementation. We implement T-MASS based on X-Pool [6], as introduced in Section 4.1. During training, we take the text embedding $\mathbf{t}$ as an condition to compute the pooled video embedding $\mathbf{v}$, and then compute the symmetric cross entropy between the randomly sampled stochastic text embedding $\mathbf{t}_s$ and text-conditioned video embedding $\mathbf{v}$. For any text-video pairs in evaluation, we still firstly compute the text-conditioned video embedding $\mathbf{v}$ upon $\mathbf{t}$, and then randomly sample a group of stochastic text embedding $\mathbf{t}_s$. We select the $\mathbf{t}_s$ that has the highest similarity with $\mathbf{v}$. Notably, there is another choice of computing pooled video embedding $\mathbf{v}$ by feeding $\mathbf{t}_s$ in fusion module $\psi(\cdot)$. However, using $\mathbf{t}_s$ in $\psi(\cdot)$ will bring randomness to $\mathbf{v}$, leading to: 1) if we resample $\mathbf{v}$ independent to $\mathbf{t}_s$ sampling for inference, the computational cost will further increase by an order; 2) if varying $\mathbf{v}$ upon $\mathbf{t}_s$, we observe a degraded performance on MSRVTT9K. Overall, we adopt $\mathbf{t}$ in $\psi(\cdot)$ following the original design of X-Pool.

Despite that the proposed method can introduce more computational cost for selecting the best stochastic text embedding $\mathbf{t}_s$ during evaluation, the complexity is a linear function of the sampling trials, *i.e.*, $\mathcal{O}(M)$, which is easy to scale and is negligible when M is small. Correspondingly, the inference speed can be comparable to existing meth-

*The work does not relate to author's position at Amazon.

†Corresponding authors: Dongfang Liu (dongfang.liu@rit.edu) and Zhiqiang Tao (zhiqiang.tao@rit.edu)

Dataset	DSL [4]	R@1	R@5	R@10	MdR	MnR
MSRVTT [14]	✗	50.2	75.3	85.1	1.0	11.9
	✓	**52.7**	**80.3**	**87.3**	**1.0**	**10.0**
LSMDC [11]	✗	28.9	48.2	57.6	6.0	43.3
	✓	**30.5**	**51.4**	**60.6**	**5.0**	**40.6**
DiDeMo [1]	✗	50.9	77.2	85.3	1.0	12.1
	✓	**55.0**	**80.9**	**87.5**	**1.0**	**9.7**
VATEX [13]	✗	63.0	92.3	96.4	1.0	3.2
	✓	**70.6**	**94.6**	**97.7**	**1.0**	**2.4**
Charades [12]	✗	14.2	36.2	48.3	12.0	54.8
	✓	**16.6**	**40.3**	**51.5**	**10.0**	**45.1**

Table 1. Text-to-video performance of T-MASS with the backbone of CLIP-ViT-B/32. "DSL" is the post processing operation.

ods by perform sampling in parallel. Since the proposed method does not introduce large components such as diffusion model [7] or temporal modeling [2, 8, 10], T-MASS is easy to be implemented, deployed, and trained with limited GPUs, such as a single NVIDIA A100 or A6000 GPU.

Similarity-aware Radius. In Fig. 3∼10, we provide more visualizations about the dynamics of $\mathcal{R}$. Specifically, we plot the $|\mathcal{R}|_1$ between one specific query text and 1000 video candidates in MSRVTT-1K. For the relevant text-video pairs, T-MASS can learn a very specific semantics range ($|\mathcal{R}|_1$ is smallest), as shown by Fig. 3∼9. In Fig. 10, we also provide an failing case where T-MASS fails to accurately capture the semantics range given the relevant text-video pair. We also visualize the radius $\mathcal{R} \in \mathbb{R}^{512}$ for both relevant pair (red) and irrelevant pairs in Fig. 2. Some dimensions may have a different scale from others. We hope our observation can bring insights toward the future design.

Stochastic Embedding with Unlearnable Radius. We provide more analysis on the proposed stochastic text embedding with unlearnable radius design, *i.e.*, $\mathcal{R} = \exp(\frac{\sum S_i}{T'})$. Notably, T-MASS with unlearnable $\mathcal{R}$ is not an inference-time method since $\mathbf{t}_s$ and $\mathcal{R}$ take effect through reparameterization using $\mathcal{L}_s$. To explore sampling t_s without learning, we implement a new baseline $\mathcal{R}$ w/o $\mathcal{L}_s$ that directly adopts the sampling. Table 3 compares between without and with $\mathcal{L}_s$, indicating the effectiveness of $\mathcal{L}_s$.

Notably, the baseline of "$\mathcal{R}$ w/o $\mathcal{L}_s$" is sampling in the original embedding space of X-Pool (since we still adopt original $\mathcal{L}_{\text{ce}}$ and there is no additional learnable module introduced). As shown in Table 3, $\mathcal{R}$ w/o $\mathcal{L}_s$ underperforms X-Pool, since the X-Pool's original space is only learned for point-to-point matching – there's no guarantee that samples around $\mathbf{t}$ could be closer to $\mathbf{v}$. Also, we observe increasing M in the original space has less effect.

3. Text Length Analysis

We analysis the effect of the text length toward the performance of proposed T-MASS. Specifically, we augment texts with the public captioner LLaVA [9] on MSRVTT-9K (4 frames taken per video) to increase the text token length from $6.91M$ to $11.52M$. We retrain X-Pool and T-MASS on the augmented dataset. The {total text token length *v.s.* R@1-gain} is $\{11.52M$ *v.s.* $+3.0\}$, compared with $\{6.91M$ *v.s.* $+3.3\}$ in Table 2 in manuscript. T-MASS can improve X-Pool (+3.0 by R@1) on the augmented text.

Dataset	DSL [4]	R@1	R@5	R@10	MdR	MnR
MSRVTT [14]	✗	52.7	77.1	85.6	1.0	10.5
	✓	**55.9**	**80.4**	**89.6**	**1.0**	**9.7**
LSMDC [11]	✗	30.3	52.2	61.3	5.0	40.1
	✓	**31.5**	**53.9**	**63.0**	**4.0**	**36.6**
DiDeMo [1]	✗	53.3	80.1	87.7	1.0	9.8
	✓	**58.0**	**82.8**	**88.9**	**1.0**	**7.5**
VATEX [13]	✗	65.6	93.9	97.2	1.0	2.7
	✓	**73.6**	**95.7**	**98.1**	**1.0**	**2.1**
Charades [12]	✗	26.7	51.7	63.9	5.0	30.0
	✓	**30.1**	**55.1**	**67.0**	**4.0**	**24.2**

Table 2. Text-to-video performance of T-MASS with the backbone of CLIP-ViT-B/16. "DSL" is the post processing operation.

Method	M	R@1	R@5	R@10	MdR	MnR
X-Pool	–	46.9	72.8	82.2	2.0	14.3
$\mathcal{R}$ w/o $\mathcal{L}_s$	1	25.6	49.1	60.9	6.0	39.1
$\mathcal{R}$ w/o $\mathcal{L}_s$	10	37.8	65.2	77.2	3.0	18.3
$\mathcal{R}$ w/o $\mathcal{L}_s$	20	37.5	66.0	76.1	2.0	17.2
$\mathcal{R}$ w/ $\mathcal{L}_s$	20	48.7	74.7	83.7	2.0	12.7

Table 3. Sampling $\mathbf{t}_s$ using unlearnable $\mathcal{R}$ without $\mathcal{L}_s$ on MSRVTT 1K. The last line is our ablated model in Table 5.

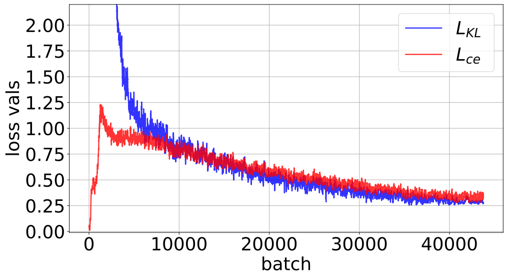

Figure 1. We explored KL-divergence, rather than adopting learnable radius design, to regularize the scale of text mass in our preliminary study. We find an unfavorable performance ($<$ 26% at R@1) even though the KL loss term shows a tendency of convergence. We adopt prior $\mathcal{N}(0, 1)$. An improper prior hardly makes contribution to the text-video alignment.

4. KL Divergence

To regularize the scale of the text mass, we explored KL divergence in our preliminary study, rather than adopting a learnable radius modeling. We find it is hard to posit a proper prior for the KL divergence term for two reasons: (1) the scale of the text mass may vary for different input texts. (2) The scale of the text mass can vary for different dimensions in the embedding space. Due to these, we find KL divergence regularization hardly contributes to the alignment of text and video, yielding an unfavorable performance, *e.g.*, $<$ 26% at R@1, even though the KL loss curve has a tendency of convergence as shown in Fig. 1.

Dataset	Text-to-video					Video-to-text				
	R@1	R@5	R@10	MdR	MnR	R@1	R@5	R@10	MdR	MnR
MSRVTT [14]	50.2	75.3	85.1	1.0	11.9	50.9	80.2	88.0	1.0	7.4
LSMDC [11]	28.9	48.2	57.6	6.0	43.3	26.0	48.4	57.5	6.0	37.8
DiDeMo [1]	50.9	77.2	85.3	1.0	12.1	49.1	76.4	85.9	2.0	8.0
VATEX [13]	63.0	92.3	96.4	1.0	3.2	76.4	98.3	99.5	1.0	1.5
Charades [12]	14.2	36.2	48.3	12.0	54.8	13.2	37.3	48.5	11.0	56.1

Table 4. Text-to-video and video-to-text performance of proposed T-MASS. We adopt the backbone of CLIP-ViT-B/32.

Dataset	Batch size	R@1	R@5	R@10	MdR	MnR
LSMDC [11]	32	28.9	48.2	57.6	6.0	43.3
	64	**29.4**	**50.4**	**58.9**	**5.0**	**41.8**
DiDeMo [1]	32	50.9	77.2	85.3	1.0	12.1
	64	**52.0**	**79.1**	**86.6**	**1.0**	**10.8**

Table 5. Text-to-video performance of proposed T-MASS with the backbone of CLIP-ViT-B/32 with larger setting.

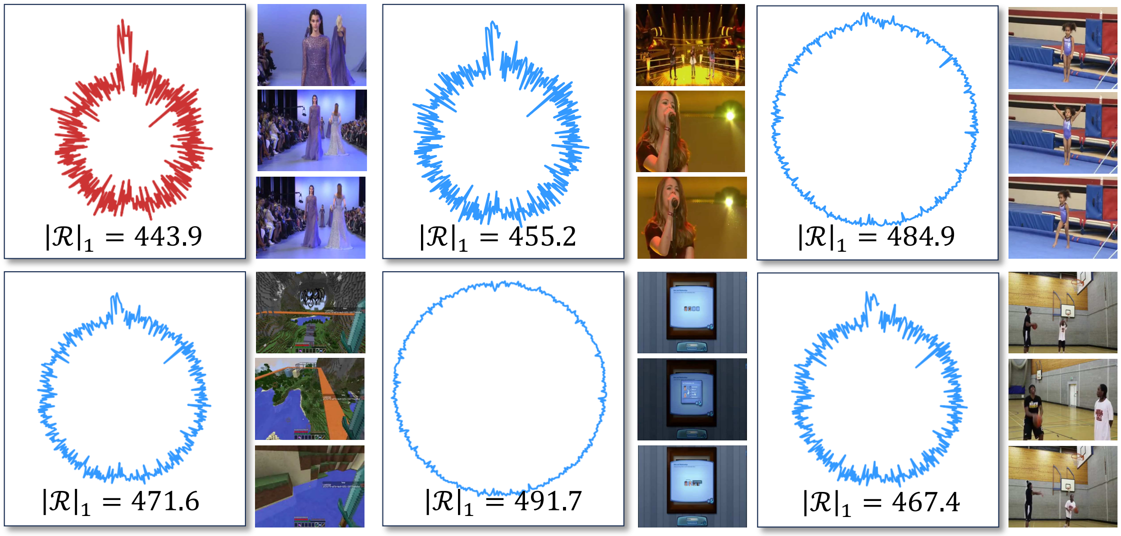

Figure 2. Visualization of $\mathcal{R} \in \mathbb{R}^{512}$. Given the text query "women are modeling clothes". We visualize the radius $\mathcal{R}$ of the relevant video in red and provide five representative radius of irrelevant videos in blue. The L_1-norm of $\mathcal{R}$ for the relevant video is smallest, corresponding to a specific text semantic range (text mass). By comparison, text mass can be larger for the irrelevant pairs.

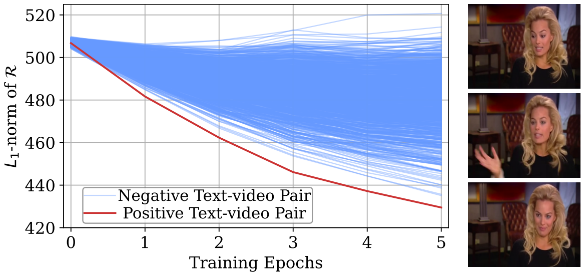

Figure 3. Dynamics of $\mathcal{R}$. We plot $|\mathcal{R}|_1$ for a relevant t-v pair (347-th in MSRVTT-1K, video on the *right*) and the query text with 999 irrelevant videos.

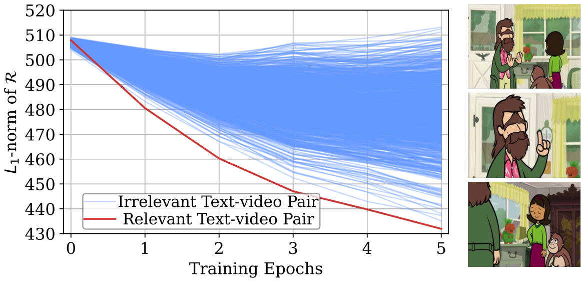

Figure 4. Dynamics of $\mathcal{R}$. We plot $|\mathcal{R}|_1$ for a relevant t-v pair (823-th in MSRVTT-1K, video on the *right*) and the query text with 999 irrelevant videos.

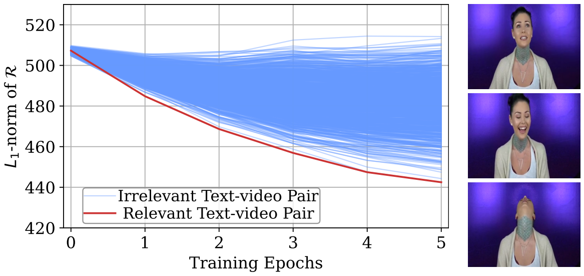

Figure 5. Dynamics of $\mathcal{R}$. We plot $|\mathcal{R}|_1$ for a relevant t-v pair (455-th in MSRVTT-1K, video on the *right*) and the query text with 999 irrelevant videos.

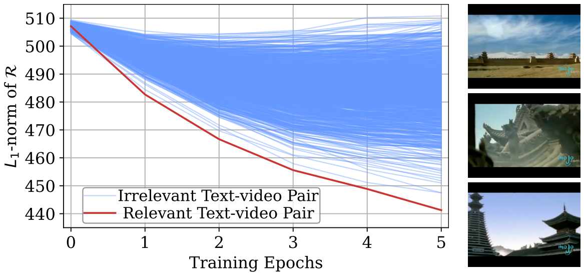

Figure 6. Dynamics of $\mathcal{R}$. We plot $|\mathcal{R}|_1$ for a relevant t-v pair (565-th in MSRVTT-1K, video on the *right*) and the query text with 999 irrelevant videos.

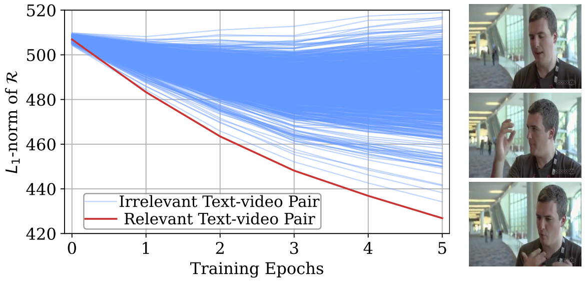

Figure 7. Dynamics of $\mathcal{R}$. We plot $|\mathcal{R}|_1$ for a relevant t-v pair (740-th in MSRVTT-1K, video on the *right*) and the query text with 999 irrelevant videos.

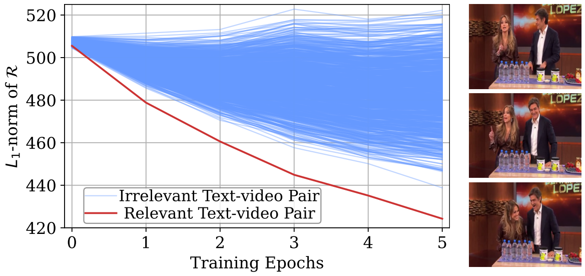

Figure 8. Dynamics of $\mathcal{R}$. We plot $|\mathcal{R}|_1$ for a relevant t-v pair (522-th in MSRVTT-1K, video on the *right*) and the query text with 999 irrelevant videos.

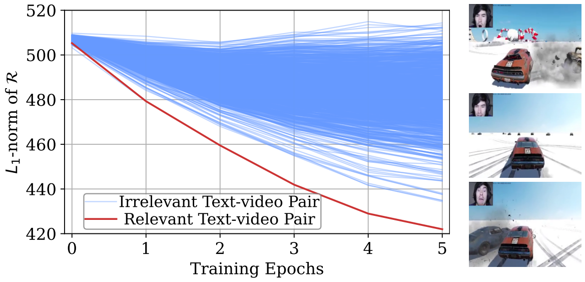

Figure 9. Dynamics of $\mathcal{R}$. We plot $|\mathcal{R}|_1$ for a relevant t-v pair (98-th in MSRVTT-1K, video on the *right*) and the query text with 999 irrelevant videos.

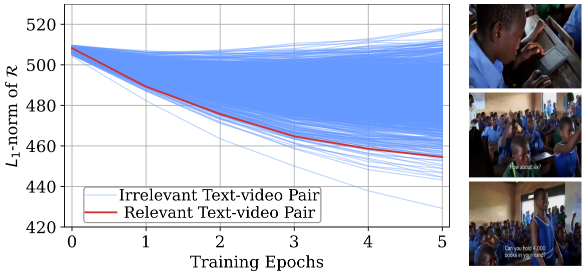

Figure 10. Dynamics of $\mathcal{R}$. We plot $|\mathcal{R}|_1$ for a relevant t-v pair (666-th in MSRVTT-1K, video on the *right*) and the query text with 999 irrelevant videos.

References

[1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *ICCV*, 2017. 1, 2, 3

[2] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, 2021. 2

[3] Simion-Vlad Bogolin, Ioana Croitoru, Hailin Jin, Yang Liu, and Samuel Albanie. Cross modal retrieval with querybank normalisation. In *CVPR*, 2022. 1

[4] Xing Cheng, Hezheng Lin, Xiangyu Wu, Fan Yang, and Dong Shen. Improving video-text retrieval by multi-stream corpus alignment and dual softmax loss. *arXiv*, 2021. 1, 2

[5] Bo Fang, Chang Liu, Yu Zhou, Min Yang, Yuxin Song, Fu Li, Weiping Wang, Xiangyang Ji, Wanli Ouyang, et al. Uatvr: Uncertainty-adaptive text-video retrieval. In *ICCV*, 2023. 1

[6] Satya Krishna Gorti, Noël Vouitsis, Junwei Ma, Keyvan Golestan, Maksims Volkovs, Animesh Garg, and Guangwei Yu. X-pool: Cross-modal language-video attention for text-video retrieval. In *CVPR*, 2022. 1

[7] Peng Jin, Hao Li, Zesen Cheng, Kehan Li, Xiangyang Ji, Chang Liu, Li Yuan, and Jie Chen. Diffusionret: Generative text-video retrieval with diffusion model. In *ICCV*, 2023. 2

[8] Y. Li, K. Min, S. Tripathi, and N. Vasconcelos. Svitt: Temporal learning of sparse video-text transformers. In *CVPR*, 2023. 2

[9] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2024. 2

[10] Yuqi Liu, Pengfei Xiong, Luhui Xu, Shengming Cao, and Qin Jin. Ts2-net: Token shift and selection transformer for text-video retrieval. In *ECCV*, 2022. 2

[11] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. A dataset for movie description. In *CVPR*, 2015. 2, 3

[12] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV*, 2016. 1, 2, 3

[13] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *ICCV*, 2019. 1, 2, 3

[14] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, 2016. 1, 2, 3

[15] Hongwei Xue, Yuchong Sun, Bei Liu, Jianlong Fu, Ruihua Song, Houqiang Li, and Jiebo Luo. Clip-vip: Adapting pre-trained image-text model to video-language representation alignment. In *ICLR*, 2023. 1