Supplementary Material for
Unified Coarse-to-Fine Alignment for Video-text Retrieval

Ziyang Wang, Yi-Lin Sung, Feng Cheng, Gedas Bertasius, Mohit Bansal
UNC Chapel Hill
{ziyangw, ylsung, fengchan, gedas, mbansal}.cs.unc.edu

Our supplementary material consists of:

(1) Additional Quantitative Results

(2) Additional Qualitative Results

(3) Method Details

1. Additional Quantitative Results

In this section, we report additional quantitative results for our UCoFiA model. First, we report the results with full video-text retrieval metrics (including video-to-text retrieval) on MSR-VTT, ActivityNet, and DiDeMo. Results indicate our UCoFiA model achieves better results on both text-to-video and video-to-text retrieval compared to the current state-of-the-art CLIP-based approaches. Meanwhile, we show UCoFiA is capable of adapting to other advanced backbone model. Then, we compare the performance and computational cost of UCoFiA with previous work and validate our methods accomplish significant improvement with limited additional computation. Lastly, we ablate the different training settings for encoders and the model design of our bi-directional ISA module (Bi-ISA).

1.1. Results with Full Metrics

In this section, we report the video-text retrieval results on MSR-VTT [19], ActivityNet [9] and DiDeMo [1] with full video-text retrieval metrics, including the results on video-to-text retrieval setting.

MSR-VTT. As shown in Table 9, UCoFiA achieves state-of-the-art results on most metrics. Specifically, compared to the most recent multi-level alignment method X-CLIP [15], UCoFiA achieves a 3.3% gain on text-to-video R@1 metric and obtains comparable results on video-to-text retrieval metrics. Compared to another recent state-of-the-art CLIP-based method TS2-Net [13], our model gets 2.4% and 1.8% improvement on R@1 metric for text-to-video and video-to-text retrieval. These results verify the effectiveness of the UCoFiA model. Moreover, replacing the visual backbone (ViT-32) with a larger model (ViT-16) would improve the model performance, especially on video-to-text retrieval.

ActivityNet. As shown in Table 10, UCoFiA outperforms the current state-of-the-art CLIP-based methods on a wide range of metrics on ActivityNet benchmark [9]. Concretely, our model achieves 1.4% and 2.4% gain on the R@1 metric on text-to-video and video-to-text retrieval compared to the state-of-the-art approaches. This indicates our UCoFiA model is capable of tackling long video retrieval, thus validating the generalization ability of our method.

DiDeMo. As shown in Table 11, compared to the current state-of-the-art models, UCoFiA achieves better results on most evaluation metrics. Specifically, our model outperforms the recent state-of-the-art CLIP-based approach X-CLIP [15] with a significant margin of 1.3% on text-to-video R@1 and 2.9% on video-to-text R@1.

1.2. Adapt UCoFiA to Other Backbone Model

In this section, we apply UCoFiA to the recent CLIP-ViP's [20] backbone model, which is a video-text model pretrained on 100M video-text pairs. As shown in Table 12, UCoFiA improves the backbone CLIP-ViP model on all metrics on the MSR-VTT text-to-video retrieval task. This indicates that our method is able to generalize to a more advanced backbone model and verifies the robustness of our method.

1.3. The Computational Cost of UCoFiA

In this section, we compare our model with the recent X-CLIP model [15] on the balance of model performance and computational cost in Table 13. Results show that UCoFiA is **3.3%** better than X-CLIP on text-to-video retrieval on MSR-VTT dataset while only requiring **1.2%** additional parameters and **1.4** GB memory per GPU (train on 4 GPUs). Therefore, UCoFiA achieves significant improvement with limited additional computational cost compared to previous works.

Method	Text → Video					Video → Text				
	R@1	R@5	R@10	MdR↓	MnR↓	R@1	R@5	R@10	MdR↓	MnR↓
CE [12]	20.9	48.8	62.4	6.0	28.2	20.6	50.3	64.0	5.3	25.1
MMT [7]	26.6	57.1	69.6	4.0	24.0	27.0	57.5	69.7	3.7	21.3
Support set [17]	27.4	56.3	67.7	3.0	-	26.6	55.1	67.5	3.0	-
Frozen [2]	31.0	59.5	70.5	3.0	-	-	-	-	-	-
HiT [11]	30.7	60.9	73.2	2.6	-	32.1	62.7	74.1	3.0	-
TT-CE [5]	29.6	61.6	74.2	3.0	-	32.1	62.7	75.0	3.0	-
CLIP-straight [18]	31.2	53.7	64.2	4.0	-	27.2	51.7	62.6	5.0	-
CLIP4Clip [14]	44.5	71.4	81.6	**2.0**	15.3	42.7	70.9	80.6	**2.0**	11.6
CAMoE [4]	44.6	72.6	81.8	**2.0**	13.3	45.1	72.4	83.1	**2.0**	10.0
X-pool [8]	46.9	72.8	82.2	**2.0**	14.3	-	-	-	-	-
X-CLIP [15]	46.1	73.0	83.1	**2.0**	13.2	46.8	73.3	**84.0**	**2.0**	**9.1**
TS2-Net [13]	47.0	**74.5**	**83.8**	**2.0**	13.0	45.3	74.1	83.7	**2.0**	9.2
UCoFiA(ViT-32)	**49.4**	72.1	83.5	**2.0**	**12.9**	**47.1**	**74.3**	83.0	**2.0**	11.4
UCoFiA(ViT-16)	49.8	74.6	83.5	2.0	13.3	49.1	77.0	83.8	2.0	11.2

Table 9. Comparison to the state-of-the-art video-text retrieval methods on MSR-VTT. The top section shows the results of non-CLIP methods and the middle section shows the results of CLIP-based methods. The bottom section shows the UCoFiA performance on different size of backbone. For fair comparison, we highlight the best results of each metric using the same backbone model (ViT-32).

Method	Text → Video					Video → Text				
	R@1	R@5	R@10	MdR↓	MnR↓	R@1	R@5	R@10	MdR↓	MnR↓
CE [12]	18.2	47.7	91.4	6.0	23.1	17.7	46.6	-	6.0	24.4
MMT [7]	28.7	61.4	94.5	3.3	16.0	28.9	61.1	-	4.0	17.1
Support set [17]	29.2	61.6	94.7	3.0	-	28.7	60.8	-	**2.0**	-
TT-CE [5]	23.5	57.2	96.1	4.0	-	23.0	56.1	-	4.0	-
CLIP4Clip [14]	40.5	72.4	**98.2**	**2.0**	7.5	42.5	74.1	85.8	**2.0**	**6.6**
TS2-Net [13]	41.0	73.6	84.5	**2.0**	8.4	-	-	-	-	-
X-CLIP [15]	44.3	74.1	-	-	7.9	43.9	73.9	-	-	7.6
UCoFiA(ours)	**45.7**	**76.6**	86.6	**2.0**	**6.4**	**46.3**	**76.5**	**86.3**	**2.0**	6.7

Table 10. Video-text retrieval results on ActivityNet.

1.4. Different Training Settings for Encoders

In this section, we show the performance of UCoFiA under different training settings for encoders. Our approach attains 47.1%, 49.4%, 43.8% R@1 on text-to-video retrieval on MSR-VTT dataset when using 1e−6, 1e−7 and 0 (frozen) learning rates for the encoders, respectively. We conjecture that CLIP is pretrained on very large-scale image-text pairs (400M) and thus we only need to slightly adjust its parameters for downstream tasks, which is also observed by many previous works (CLIP4Clip, X-CLIP).

1.5. Additional Ablation Study for Bi-ISA

In the main paper, we mention that empirically we find that jointly considering two directions of patch-word matrix aggregation (patch-then-word and word-then-patch) provides better aggregation for the patch-word matrix. In Table 14, we compare our bi-directional solution with single-directional methods on MSR-VTT dataset. For better comparison, we do not apply the Sinkhorn Knopp algorithm to normalize the retrieval similarities. Results show that leveraging both aggregation directions achieves better results, validating the effectiveness of our Bi-ISA design.

2. Additional Qualitative Results

In this section, we provide additional qualitative results of UCoFiA. First, we visualize the imbalanced retrieval results and show how our unification module mitigates this issue. Then, we visualize the video samples retrieved by methods focusing on different alignment levels to validate the effectiveness of our coarse-to-fine alignment design.

2.1. Visualization of Imbalanced Retrieval

As discussed in the main paper, we find that scores across different videos are highly imbalanced in the similarity matrices of each level. As a result, the video candidate could be over-/under-represented by the retrieval model due to the imbalanced summation of retrieval similarities. As shown in Figure 5, the left part denotes the video candidates haven't been retrieved in the inference stage which corresponds to under-representative. The right part de-

Method	Text → Video					Video → Text				
	R@1	R@5	R@10	MdR↓	MnR↓	R@1	R@5	R@10	MdR↓	MnR↓
CE [12]	16.1	41.1	-	8.3	43.7	15.6	40.9	-	8.2	42.4
ClipBERT [10]	20.4	48.0	60.8	6.0	-	-	-	-	-	-
TT-CE [5]	34.6	65.0	74.7	3.0	-	21.1	47.3	61.1	6.3	-
Frozen [2]	21.6	48.6	62.9	6.0	-	-	-	-	-	-
CLIP4Clip [14]	43.4	70.2	80.6	**2.0**	17.5	42.5	70.6	80.2	**2.0**	11.6
TS2-Net [13]	41.8	71.6	82.0	**2.0**	14.8	-	-	-	-	-
X-CLIP [15]	45.2	74.0	-	-	14.6	43.1	**72.2**	-	-	**10.9**
UCoFiA(ours)	**46.5**	**74.8**	**84.4**	**2.0**	**13.4**	**46.0**	71.9	**81.5**	**2.0**	12.1

Table 11. Video-text retrieval results on DiDeMo.

Methods	R@1	R@5	R@10
CLIP-ViP [20]	50.1	74.8	84.6
UCoFiA with CLIP-ViP	**51.3**	**75.1**	**85.2**

Table 12. Text-to-video retrieval results on MSR-VTT dataset under CLIP-ViP backbone model.

Model	R@1	Param (M) ↓	Mem (GB) ↓
X-CLIP [15]	46.1	**164**	**12.5**
UCoFiA(ours)	**49.4**	166	13.9

Table 13. The comparison of performance (text-to-video retrieval on MSR-VTT) and computational cost (model parameters and memory) between X-CLIP and UCoFiA(ours).

Patch-then-word	Word-then-patch	R@1	R@5	MnR↓
✓		47.8	72.8	13.4
	✓	47.7	72.9	13.5
✓	✓	**48.2**	**73.3**	**13.2**

Table 14. The effect of leveraging both aggregation directions (patch-then-word and word-then-patch) for patch-word matrix aggregation on MSR-VTT dataset. The last row is our design.

notes the video candidates have been retrieved more than twice (including twice) in the inference stage which corresponds to over-representative. The middle part denotes the video candidates have been retrieved once, which is the ideal situation. The blue column in Figure 5 represents the model without the Sinkhorn Knopp algorithm. The results show that only 43% video candidates are retrieved once in the inference stage while 34% video candidates are under-represented and 23% video candidates are over-represented. After applying the Sinkhorn Knopp algorithm in the unification module (the orange column in Figure 5), the under-representative issue is mitigated and more than 50 under-represented video candidates have been re-scaled and retrieved by the model. Meanwhile, we also observe a slight reduction in the number of over-represented videos. In all, the Sinkhorn Knopp algorithm in the unification module indeed mitigates the over- and under-representation issue in the inference stage.

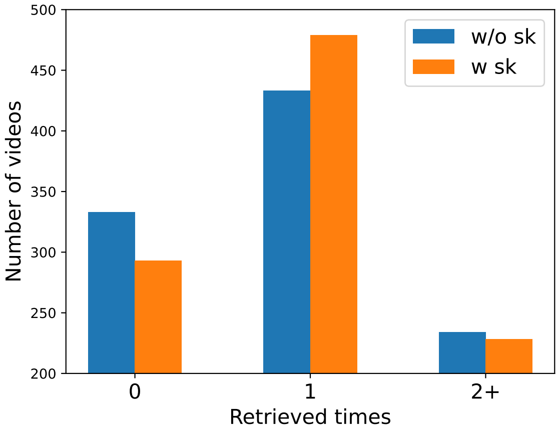

Figure 5. The visualization of imbalanced retrieval results. The left part denotes the video candidates haven't been retrieved in the inference stage (under-representative). The middle part denotes the video candidates have been retrieved once in the inference stage. The right part denotes the video candidates have been retrieved more than twice (including twice) in the inference stage (over-representative). The blue column represents the model without the Sinkhorn Knopp algorithm and the orange column represents the full UCoFiA model. Results show that about 50 under-represented videos have been re-scaled and retrieved by the videos via the SK algorithm (from the left bar to the middle bar).

2.2. Comparison of Different Alignments

As discussed in the main paper, our coarse-to-fine alignment module captures comprehensive cross-modal clues compared to coarse-grained or fine-grained alignment. We provide more visualization results in Figure 7. For the first text query (on the first row of Figure 7), the coarse-grained alignment only captures the scene of "singing" and the fine-grained alignment only focuses on the object "guitar". For the second text query (on the second row of Figure 7), the coarse-grained alignment only considers the scene information like "driving", and "video game" while the fine-grained alignment only captures the detail information "motorcycle". For the last text query (on the last row of Figure 7),

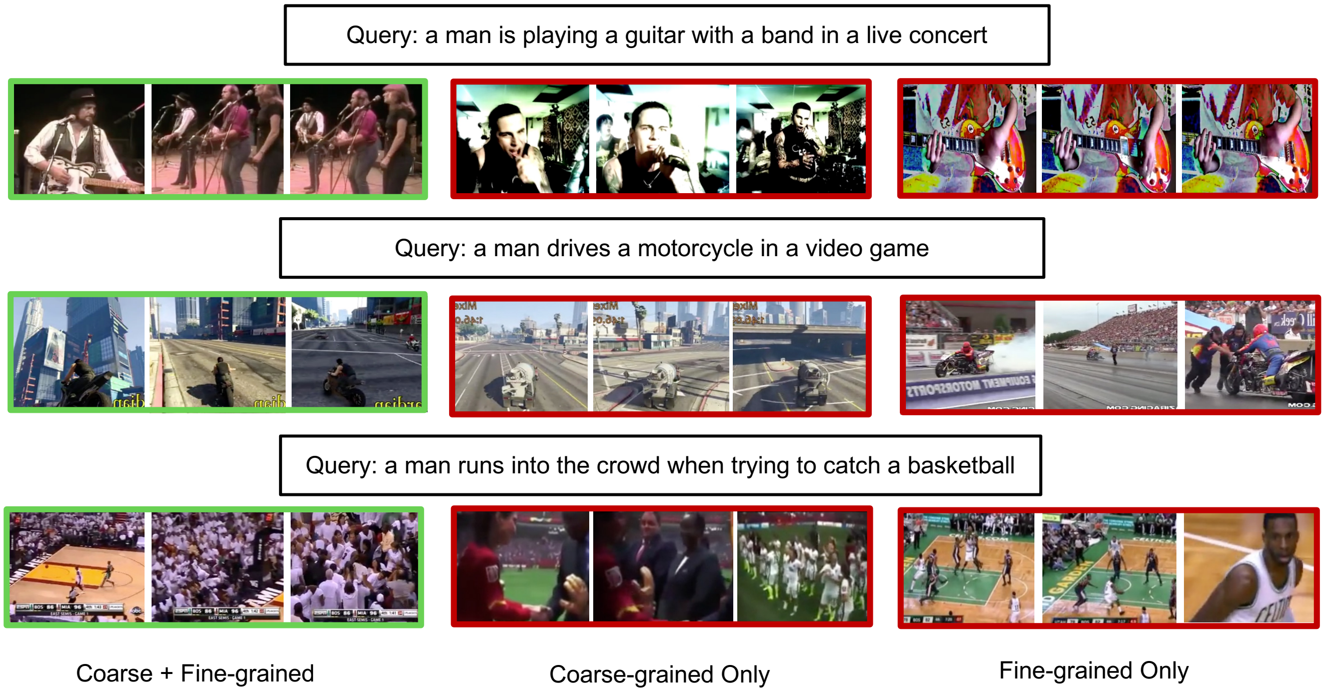
Query: a man is playing a guitar with a band in a live concert
Query: a man drives a motorcycle in a video game
Query: a man runs into the crowd when trying to catch a basketball
Coarse + Fine-grained
Coarse-grained Only
Fine-grained Only

Figure 6. The visualization of different alignments. The left part is the correctly retrieved video by our coarse-to-fine alignment module. The middle part is the wrongly retrieved video by coarse-grained alignment and the right part is the wrongly retrieved video by fine-grained alignment.

Figure 7. The visualization of different alignments. The left part is the correctly retrieved video by our coarse-to-fine alignment module. The middle part is the wrongly retrieved video by coarse-grained alignment and the right part is the wrongly retrieved video by fine-grained alignment.

the coarse-grained alignment overlooks the detailed information "basketball" and the fine-grained alignment ignores the scene of "crowd" and the action of "run into". To sum up, the coarse-grained or fine-grained alignment could overlook some crucial cross-modal clues while our coarse-to-fine alignment is capable of capturing both high-level and detailed information and retrieving the correct video candidate.

3. Method Details

In this section, we present more details of UCoFiA. First, we discuss the patch selection module. Then, we present details of the Sinkhorn-Knopp Algorithm that normalizes the similarity matrix for unification.

3.1. Patch Selection Module

As discussed in the main paper, due to the high redundancy of patch tokens, inspired by [13], we propose a patch selection module to choose the top-K salient patches from each frame for patch-word alignment. Here we present the details of the patch selection module.

Specifically, given the patch feature for the n-th frame p_n, where $p_n = \mathcal{F}_v\left(F_n\right) \in \mathbb{R}^{M \times C}$, M denotes the number of the visual patches within a video, we select the top-K salient token out of the M tokens of the frame. To allow each patch to be aware of the information of the whole frame, we first concatenate the frame feature $f \in \mathbb{R}^C$ with each patch feature and leverage an MLP layer to fuse the global (frame) and local (patch) information, and leverage an MLP layer $\mathcal{G}_a$ to obtain the frame-augmented patch information to mitigate the influence of irrelevant background patches. Then, to avoid the selection module only considering the frame information and deviating from the information of the original video, we further concatenate the frame-augmented patch information with the video representation v and apply another MLP layer $\mathcal{G}_b$ to obtain a saliency score U for each patch. The whole process can be denoted as:

$$U = \mathcal{G}_b\left(\text{Concat}\left(\mathcal{G}_a\left(\text{Concat}\left(p_n, f\right)\right), v\right)\right). \qquad (1)$$

Then, according to the saliency score U, we select the indices of K most salient patches within a video frame $ind \in \{0,1\}^K$. Through this one-hot vector ind, we extract the top-K salient patch by

$$\hat{p} = ind^T p, \qquad (2)$$

where $\hat{p}_n \in \mathbb{R}^{K \times C}$ denotes the selected patch representa-

Algorithm 1 Sinkhorn-Knopp algorithm

```
function SINKHORN-KNOPP($\mathbf{S}$, $n_{iter}$)
    $L = \mathbf{S}.\exp()$
    $\beta = 1 \,/\, L.\text{sum}(\texttt{dim} = 0)$
    for $i$ in range($n_{iter}$) do
        $\alpha = 1 \,/\, (L \; @ \; \beta)$
        $\beta = 1 \,/\, (\alpha \; @ \; L)$
    end for
    $\alpha \leftarrow \alpha.\log()$
    return $\alpha$
end function
```

tion for the whole video. We concatenate the selected patch feature from all N frames and obtain the selected patch feature $\hat{p} \in \mathbb{R}^{L_v \times C}$, where $L_v = N * K$. Note that the direct top-K patch selection is non-differentiable, in practice, to make the patch selection module differentiable, we apply the perturbed maximum method proposed in [3].

3.2. Sinkhorn-Knopp Algorithm

As discussed in the main paper, inspired by [16], we utilize the Sinkhorn-Knopp algorithm [6] to normalize the similarity scores for each granularity and make sure the marginal similarities (the sum of retrieval similarities between one specific video and all texts) for different videos are almost identical, so that each video has a fair chance to be selected. Below, we discuss the algorithm in detail.

Recall that our goal is to compute the video bias using the testing video set (G videos) and the training text set (H queries). Given the similarity matrix $\mathbf{S} \in \mathbb{R}^{G \times H}$, we leverage the Algorithm 1 to compute the video bias $\alpha \in \mathbb{R}^{G}$ in an iterative manner (the number of iterations $n_{iter} = 4$ for all datasets). The fixed-point iteration process allows the model to find the optimal value of α with minimum cost. We further add the α to the similarity logits to re-scale the similarity matrix to normalize the marginal similarity of every video to be a similar value.

References

[1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812, 2017. 1

[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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738, 2021. 2, 3

[3] Quentin Berthet, Mathieu Blondel, Olivier Teboul, Marco Cuturi, Jean-Philippe Vert, and Francis Bach. Learning with differentiable pertubed optimizers. *Advances in neural information processing systems*, 33:9508–9519, 2020. 5

[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 preprint arXiv:2109.04290*, 2021. 2

[5] Ioana Croitoru, Simion-Vlad Bogolin, Marius Leordeanu, Hailin Jin, Andrew Zisserman, Samuel Albanie, and Yang Liu. Teachtext: Crossmodal generalized distillation for text-video retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11583–11593, 2021. 2, 3

[6] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013. 5

[7] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 214–229. Springer, 2020. 2

[8] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5006–5015, 2022. 2

[9] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715, 2017. 1

[10] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7331–7341, 2021. 3

[11] Song Liu, Haoqi Fan, Shengsheng Qian, Yiru Chen, Wenkui Ding, and Zhongyuan Wang. Hit: Hierarchical transformer with momentum contrast for video-text retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11915–11925, 2021. 2

[12] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use what you have: Video retrieval using representations from collaborative experts. *arXiv preprint arXiv:1907.13487*, 2019. 2, 3

[13] Yuqi Liu, Pengfei Xiong, Luhui Xu, Shengming Cao, and Qin Jin. Ts2-net: Token shift and selection transformer for text-video retrieval. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIV*, pages 319–335. Springer, 2022. 1, 2, 3, 4

[14] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304, 2022. 2, 3

[15] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 638–647, 2022. 1, 2, 3

[16] Yookoon Park, Mahmoud Azab, Bo Xiong, Seungwhan Moon, Florian Metze, Gourab Kundu, and Kirmani Ahmed. Normalized contrastive learning for text-video retrieval. *arXiv preprint arXiv:2212.11790*, 2022. 5

[17] Mandela Patrick, Po-Yao Huang, Yuki Asano, Florian Metze, Alexander Hauptmann, Joao Henriques, and Andrea Vedaldi. Support-set bottlenecks for video-text representation learning. *arXiv preprint arXiv:2010.02824*, 2020. 2

[18] Jesús Andrés Portillo-Quintero, José Carlos Ortiz-Bayliss, and Hugo Terashima-Marín. A straightforward framework for video retrieval using clip. In *Pattern Recognition: 13th Mexican Conference, MCPR 2021, Mexico City, Mexico, June 23–26, 2021, Proceedings*, pages 3–12. Springer, 2021. 2

[19] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016. 1

[20] Hongwei Xue, Yuchong Sun, Bei Liu, Jianlong Fu, Ruihua Song, Houqiang Li, and Jiebo Luo. Clip-vip: Adapting pretrained image-text model to video-language representation alignment. *arXiv preprint arXiv:2209.06430*, 2022. 1, 3