# Renmin University of China at TRECVID 2021: Searching and Describing Video

Xirong Li, Aozhu Chen, Fan Hu, Xinru Chen, Chengbo Dong, Gang Yang
MOE Key Lab of DEKE, Renmin University of China
AIMC Lab, School of Information, Renmin University of China
https://ruc-aimc-lab.github.io

## Abstract

*In this paper, we summarize our TRECVID 2021 experiments. We participated in two tasks: Ad-hoc Video Search (AVS) and Video-to-Text Description Generation (VTT). For the AVS task, we develop our solutions based on two cross-modal matching models,* i.e. *Sentence Encoder Assembly (SEA) and Multiple Encoder Assembly (MEA). MEA is a variant of SEA that performs modality-specific attention-based feature fusion on the text side and the video side, respectively. Our best AVS run is obtained by late average fusion of MEA, SEA and CLIP, scoring mean infAP of 0.343. For the VTT task, we focus on description generation. Our VTT solutions are developed the basis of the classical Bottom-Up-Top-Down (BUTD) model, with its encoder and decoder improved. Multi-Level visual representation, reinforced adversarial learning and cross-modal matching based reranking help our best submission score CIDEr_D of 28.8. The 2021 edition of the TRECVID benchmark has been a fruitful participation for the RUCMM team. Our runs are ranked at the third place for AVS and the second place for VTT Description Generation.*

## 1 Ad-hoc Video Search

### 1.1 Approach

Our solutions for the TV21 AVS task are developed based on two cross-modal matching networks. One is the Sentence Encoder Assembly (SEA) model [21], previously used in our TV20 solution [20]. The other is an improvement of SEA, which we term Multiple Encoder Assembly (MEA). SEA supports text-video matching in multiple text-encoder-specific common spaces. MEA improves over SEA in the following two aspects. First, its text encoders are expanded to include a pre-trained CLIP model [27]. Second, for a more effective fusion of diverse visual features, visual-encode-specific common space learning is performed [15].

As illustrated in Fig. 1, given $m$ sentence-level features produced by $m$ text encoders and $n$ video-level features produced by $n$ visual encoders, MEA first performs modality-specific attention-based feature fusion to produce a new combined feature per modality [15]. By pairing each of the $m+1$ text features and each of the $n+1$ video features, a number of $(m+1)\times(n+1)$ common spaces are built. By averaging the cosine similarities computed in the individual spaces, we have the final cross-modal similarity as

$$cms(t,v) := \frac{1}{(n+1)\times(m+1)} \sum_{j=1}^{n+1}\sum_{i=1}^{m+1} cos(f_{i,j}(t), f_{i,j}(v)), \quad (1)$$

where $f_{i,j}(t)$ and $f_{i,j}(v)$ indicate respectively the text and video embeddings per common space indexed by $(i,j)$.

#### 1.1.1 Choice of Visual Encoders

We adopt a number of pre-trained 2D / 3D deep visual models as visual encoders. The following seven deep visual features are used:

1. *rx101*: A 2,048-d frame-level feature, extracted by ResNeXt-101 trained on the full ImageNet set[1] [25].
2. *re152*: A 2,048-d frame-level feature, extracted by ResNet-152 trained on 1k-class ImageNet[1] [11].
3. *wsl*: A 2,048-d frame-level feature, extracted by ResNeXt-101 pre-trained on weakly labeled web images followed by fine-tuning on ImageNet[2] [24].
4. *clip*: A 512-d frame-level feature, extracted by a pre-trained CLIP[3] model (ViT-B/32) [27].
5. *c3d*: A 2,048-d segment-level feature, extracted by C3D [31] trained on Kinetics400[4].
6. *ircsn*: A 2,048-d segment-level feature, extracted by irCSN-152 [14] trained on IG-65M[5].
7. *tf*: A 768-d segment-level feature, extracted by TimeSformer [6] pre-trained on HowTo100M[6].

[1] https://github.com/xuchaoxi/video-cnn-feat
[2] https://github.com/facebookresearch/WSL-Images
[3] https://github.com/openai/CLIP
[4] https://github.com/DavideA/c3d-pytorch
[5] https://github.com/facebookresearch/VMZ/tree/master/pt
[6] https://github.com/facebookresearch/TimeSformer

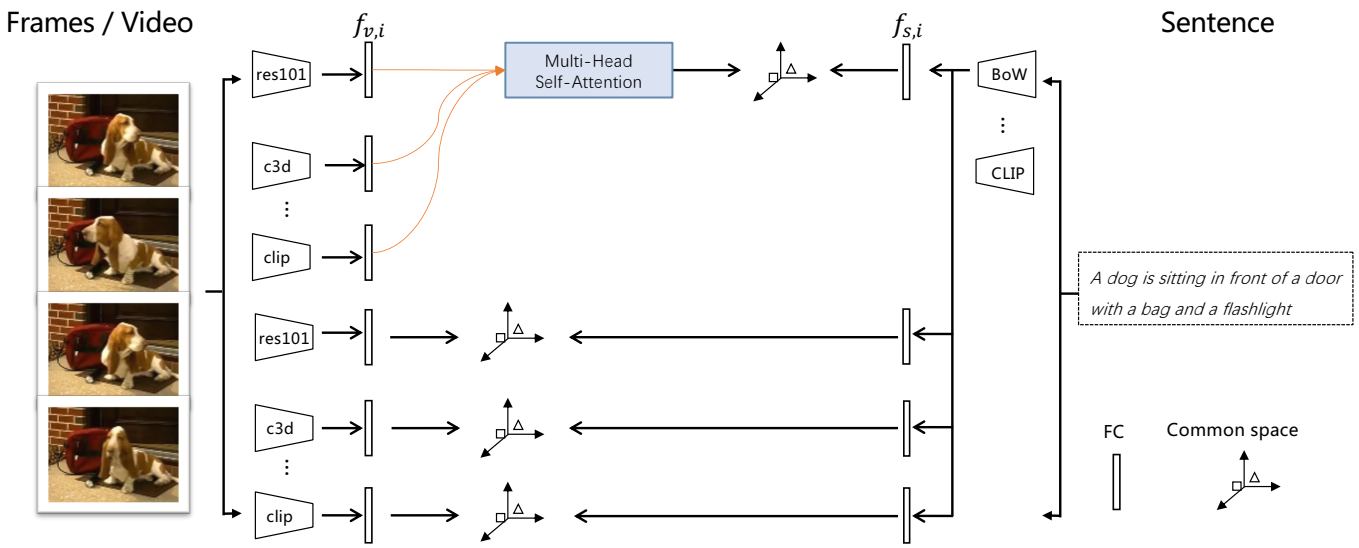


**Figure 1: Conceptual diagram of the Multiple Encoder Assembly (MEA)**. The key idea of MEA is to given $m$ sentence-level features produced by $m$ text encoders and $n$ video-level features produced by $n$ visual encoders. By pairing each of the $m+1$ text features and each of the $n+1$ video features, a number of $(m+1)\times(n+1)$ common spaces are built to compute a multi-space text-video similarity.

#### 1.1.2 Choice of Textual Encoders

We experimented with the following three sentence encoders:

1. Bag-of-Words (bow) [10],
2. word2vec (w2v) [26], pretrained on Flickr tags [10, 17],
3. CLIP (ViT-B/32) [27].

#### 1.1.3 Choice of (Pre-)Training Data

Following our earlier study on the AVS task [7], our training data consists of MSR-VTT [35], TGIF [22], and VATEX [34]. The following two image collections are used for pre-training: MS-COCO [23] and GCC [30].

### 1.2 Internal Evaluation

Following Run4 of our TV19 system [19], we use SEA(bow, w2v) for preliminary feature selection. As shown in the Table 1, Rank 3 and Rank 4 perform closely. So we further test their features, *i.e. rx101-re152-clip* and *rx101-wsl-clip*, with SEA-clip [7]. Different from SEA(bow, w2v), SEA-clip has bow, w2v and CLIP as its text encoders. The result shows that *rx101-wsl-clip* is better. We then tried to add 3D features. Table 2 shows *rx101-wsl-clip-ircsn* has the best performance. Hence, we use the following four visual features in our solutions: *rx101*, *wsl*, *clip* and *ircsn*.

For training data, we use the joint set of MSR-VTT [35], TGIF [22] and VATEX [34]. Following our conventional setup [16, 18–20], the development set of the TRECVID 2016 Video-to-Text Matching task [5] is used as an external validation set[7]. We use a two-round pre-training strategy, where a model is first pre-trained on GCC followed by another round pretraining on MS-COCO. As 3D features are unavailable on images, we simply use zero-valued vectors for pre-training.

**Table 1: Comparing 2D and 3D visual features on the TRECVID 19/20 AVS task**. Training data: MSR-VTT + TGIF.

| Rank | Visual features | TV19 | TV20 | MEAN |
|---|---|---|---|---|
| ***Retrieval model*: SEA-clip [7]** | | | | |
| 1 | *rx101-wsl-clip* | 0.204 | 0.262 | 0.233 |
| 2 | *rx101-re152-clip* | 0.199 | 0.249 | 0.224 |
| ***Retrieval model*: SEA(bow, w2v) [21]** | | | | |
| 3 | *rx101-re152-clip* | 0.199 | 0.233 | 0.216 |
| 4 | *rx101-wsl-clip* | 0.183 | 0.239 | 0.211 |
| 5 | *rx101-clip* | 0.196 | 0.223 | 0.210 |
| 6 | *rx101-re152-wsl-clip* | 0.182 | 0.238 | 0.210 |
| 7 | *rx101-re152-wsl* | 0.159 | 0.208 | 0.184 |
| 8 | *rx101-re152* | 0.167 | 0.201 | 0.184 |
| 9 | *rx101-wsl* | 0.148 | 0.208 | 0.178 |
| 10 | *clip-wsl* | 0.148 | 0.208 | 0.178 |
| 11 | *wsl* | 0.135 | 0.210 | 0.173 |
| 12 | *clip* | 0.155 | 0.180 | 0.168 |
| 13 | *ircsn* | 0.088 | 0.193 | 0.140 |
| 14 | *tf* | 0.102 | 0.142 | 0.122 |
| 15 | *c3d* | 0.036 | 0.098 | 0.067 |

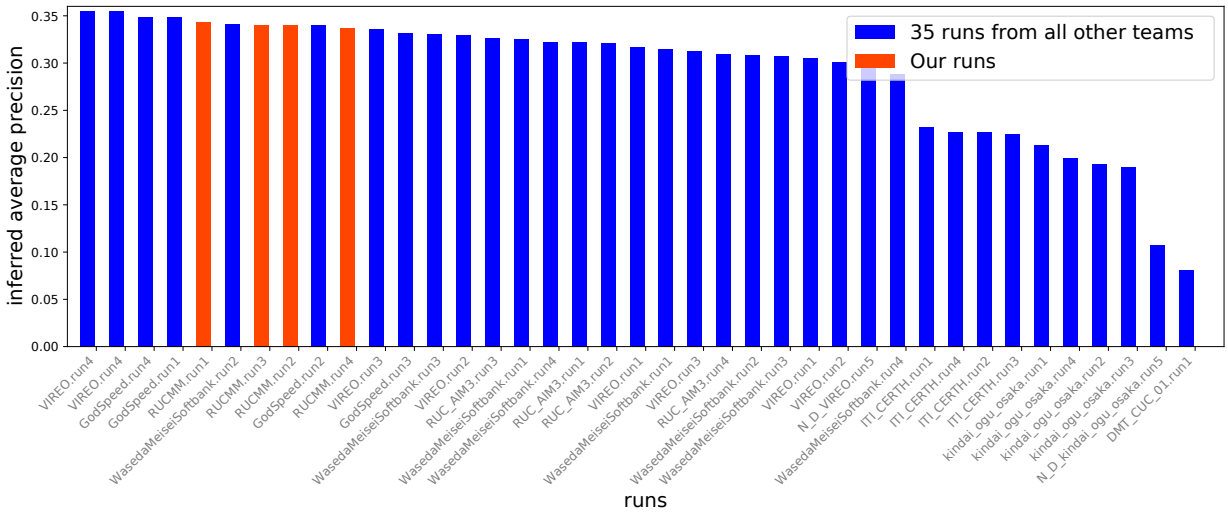


**Figure 2: Overview of the TRECVID 2021 AVS benchmark evaluation**.

**Table 2: Evaluating the influence of pre-training on the TRECVID 19/20 AVS tasks**. Retrieval model: MAE.

| Visual features | TV19 | TV20 | MEAN |
|---|---|---|---|
| ***Pre-training on MS-COCO:*** | | | |
| *rx101-wsl-clip-c3d* | 0.203 | 0.339 | 0.271 |
| *rx101-wsl-clip-tf* | 0.205 | 0.313 | 0.259 |
| *rx101-wsl-clip-ircsn* | 0.196 | 0.347 | 0.272 |
| ***Pre-training on GCC:*** | | | |
| *rx101-wsl-clip-c3d* | 0.199 | 0.331 | 0.265 |
| *rx101-wsl-clip-ircsn* | 0.203 | 0.335 | 0.269 |
| ***Pre-training on GCC and MS-COCO:*** | | | |
| *rx101-wsl-clip-ircsn* | 0.209 | 0.357 | 0.283 |

## 1.3 Submissions

Based on the performance of the individual models and their combinations on the TV18/TV19 AVS tasks, see Table 3, we submitted the following four runs:

- *Run 4*: SEA-clip
- *Run 3*: MEA
- *Run 2*: Re-ranking the results of MEA by matching concepts extracted from sentences and videos.
- *Run 1* (primary run): Late average fusion of *Run 4*, *Run 3* and CLIP.

The performance of our four runs and two baselines on the TRECVID 2021 AVS task is summarized in Table 3. Compared to the invididual models, MEA is the best, followed by SEA-clip and CLIP. An overview of the AVS task benchmark is shown in Fig. 2. Our primary run, with mean infAP of 0.343, is ranked at the third place team-wise.

[7] https://github.com/li-xirong/avs

# 2 Video-to-Text Description Generation

In this subtask, participants were asked to automatically generate a natural language sentence to describe the content of a given *unlabeled* video.

## 2.1 Approach

Our solutions are developed on the basis of the classical Bottom-Up-Top-Down (BUTD) model [1], by improving its encoder and decoder [8]. The overall architectures is illustrated in Fig. 3.

### 2.1.1 Multi-Level Video Feature Extraction and Enhancement

Multi-level visual representation learning is found to be crucial for a comprehensive representation of the video content [11, 12]. In this work, we extract both holistic scene-level features and fine-grained object-level features. Scene-level features are extracted from various pre-trained models such as ResNext [24], CLIP (ViT-B/32) [27], TimesFormer [6], X3D [13] and irCSN [14], as previously used for the AVS task. In order to extract features for salient objects, we propose to apply a pre-trained video object segmentation network, AGNN [33], instead of widely used Faster-RCNN as the object-level feature extractor. To align and enhance the scene- and object- level features, a multi-head self-attention module (MHSA) is used.

### 2.1.2 Caption Generation Network

Our caption generation network is based on the widely used Bottom-Up-Top-Down (BUTD) model [1]. In a nutshell, BUTD runs two LSTMs, *i.e.* an attention LSTM (att-LSTM) and a language LSTM (lang-LSTM), in turn. To

**Table 3: Performance of our TV21 submissions and two baselines on TRECVID 2016–2021 AVS tasks.**

| Submission | Solution | TV16 | TV17 | TV18 | TV19 | TV20 | TV21 | MEAN |
|---|---|---|---|---|---|---|---|---|
| Run1 | Late fusion (Run3, Run4, CLIP) | **0.246** | 0.320 | 0.161 | **0.227** | **0.366** | **0.343** | 0.277 |
| Run2 | MEA-rerank | 0.224 | **0.342** | **0.168** | 0.224 | 0.361 | 0.340 | **0.293** |
| Run3 | MEA | 0.223 | 0.342 | 0.167 | 0.223 | 0.361 | 0.340 | 0.292 |
| Run4 | SEA-clip | 0.232 | 0.255 | 0.135 | 0.213 | 0.358 | 0.337 | 0.239 |
| - | CLIP | 0.173 | 0.208 | 0.087 | 0.136 | 0.161 | 0.194 | 0.160 |
| - | Late fusion (Run3, Run4) | 0.235 | 0.300 | 0.156 | 0.225 | 0.365 | 0.339 | 0.270 |

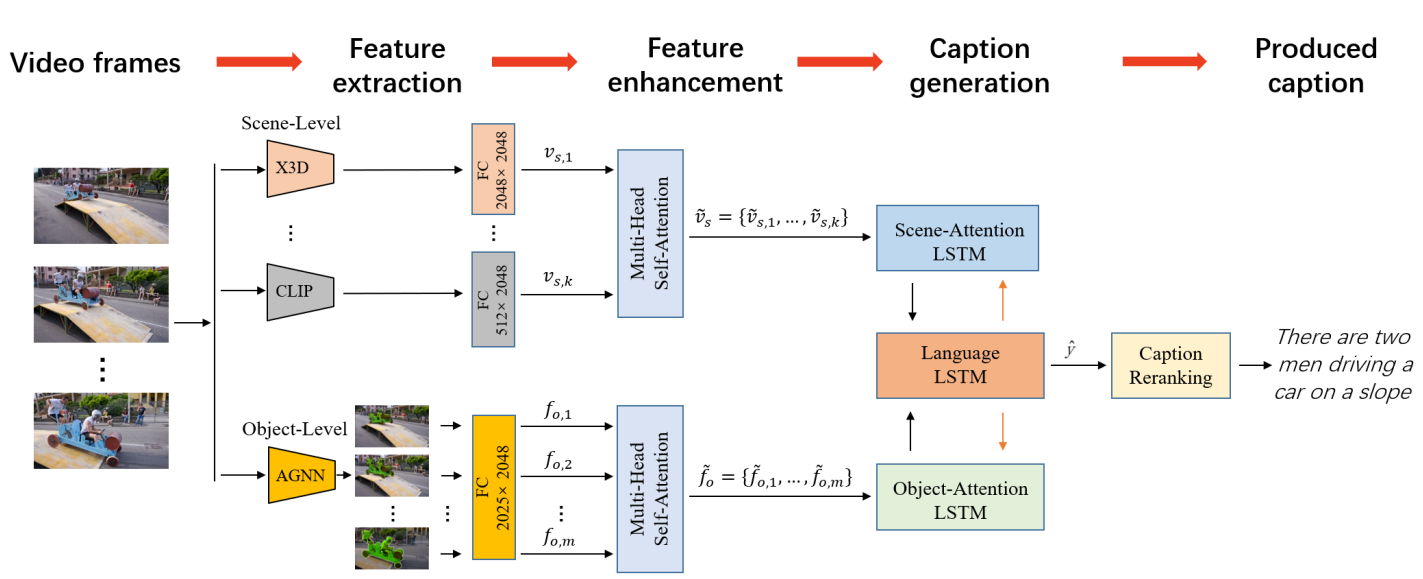


**Figure 3: Proposed solution for video captioning** [8]. The solution consists of four key steps: 1) scene-level and object-level feature extraction, 2) feature enhancement by multi-head self attention, 3) candidate caption generation by Bottom-Up-Top-Down, and 4) caption reranking by cross-modal matching. The entire network is end-to-to-end trained by reinforcement and adversarial learning, with a BERT based semantic similarity between generated and reference captions added into the reward and a jointly trained discriminator.

deal with the scene-level and object-level features, we introduce two att-LSTMs. The scene-att-LSTM concentrates on the significance of diverse scene features, while the object-att-LSTM attends the relationship among temporal frames from the object aspect. By applying such dual attention LSTMs, we effectively exploit multi-level visual representation for caption generation.

### 2.1.3 Reinforced Adversarial Learning

Towards generating human-like captions, our model is trained in a semantic-reinforced and adversarial manner besides minimizing the cross-entropy loss in word prediction. Firstly, we adopt a semantic-reinforced training strategy via self-critical sequence training (SCST), which has been proved effective in image and video captioning [1,29,36,37]. As for the choice of the rewards for reinforcement learning, previous works use metrics(*e.g.* CIDEr-D [32]). We further consider a semantic similarity-based reward. In particular, a pre-trained SentenceBert [28] is used to measure the similarity score between generated captions and corresponding ground truth. Secondly, we apply a discriminator to estimate how likely the generated captions share the same fluency and language style with the groundtruths. In such an adversarial training manner, the decoder, which can be regarded as a generator, can be optimized jointly with the discriminator towards generating more fluent and human-like captions.

### 2.1.4 Caption Reranking by Cross-Modal Matching

Multiple models are trained with different setups. In order to pick up the best caption from candidate captions generated independently by these models, a retrieval-based caption reranking is conducted [9]. To be specific, two pre-trained video-text matching models, SEA [21] and CLIP (ViT-B/32) [27], are utilized to calculate the alignment scores between the given video and each candidate caption. Both SEA and CLIP embed the video and captions into a common space, wherein the cross-modal semantic similarity is computed as the cosine similarity between the corresponding embedding vectors. We average the similarity scores from the two models. The caption with the highest score is selected.

**Table 4: Our runs in the TRECVID 2021 VTT description generation subtask.** Our best run receives CIDEr-D of 28.8%.

| Run | Feature Extraction & Enhancement | Reinforced Learning | Adversarial Learning | Reranking | Bleu_4 | METEOR | CIDEr_D | SPICE |
|---|---|---|---|---|---|---|---|---|
| 1 | ✓ | ✓ | ✗ | ✗ | 9.65 | 27.13 | 25.3 | 9.3 |
| 2 | ✓ | ✓ | ✓ | ✗ | 9.65 | 27.02 | 25.5 | 9.2 |
| 3 | ✓ | ✓ | after 5 epoch | ✗ | 9.35 | 27.05 | 25.4 | 9.3 |
| 4 | - | - | - | ✓ | **9.88** | **28.46** | **28.8** | **10.3** |

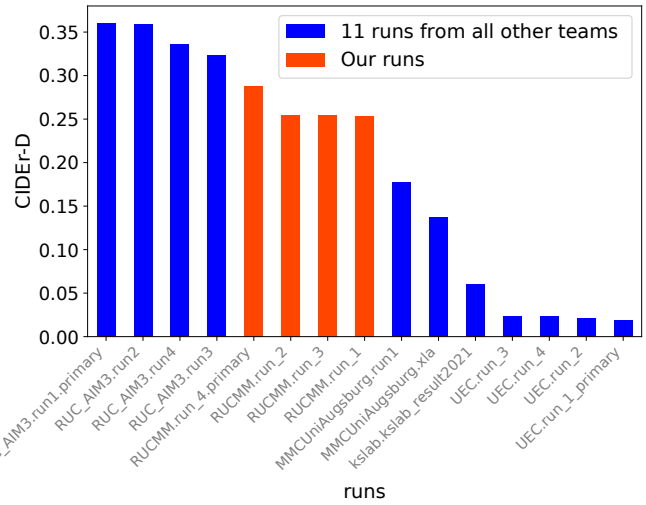


**Figure 4: Overview of the TRECVID 2021 VTT Description Generation benchmark evaluation**.

## 2.2 Submissions

We use five public datasets for training, including the training split of MSR-VTT [35], TGIF [22], VATEX [34] and TRECVID VTT 2018-2019 [2]. The VTT 2020 training set [4] is used for internal evaluation. As the effectiveness of feature extraction & enhancement, caption generation and semantic-reinforced training are separately verified in our previous work [8], we start from the internal evaluation of adversarial training. The following four runs are submitted:

- *Run 1* is the baseline model, which use merely semantic-reinforced training strategy.
- *Run 2* adds adversarial learning based on *Run 1*.
- *Run 3* starts adversarial learning after 5 epoch due to the saturation of discriminator.
- *Run 4* is the ensemble version via reranking.

Results are shown in Table 4. Adversarial training requires further exploration for better generalization, and the ensemble version via reranking shows the most significant improvements. An overview of the VTT Description Generation task benchmark is shown in Fig. 4. Team-wise, our submissions are ranked at the second place among all the submissions.

## Acknowledgments

The authors are grateful to the TRECVID coordinators for the benchmark organization effort [3]. This research was supported by the National Natural Science Foundation of China (No. 62172420, No. 61672523), Beijing Natural Science Foundation (No. 4202033), and the Fundamental Research Funds for the Central Universities and the Research Funds of Renmin University of China (No. 18XNLG19).

## References

[1] P. Anderson, X. He, C. Buehler, D. Teney, and L. Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018.

[2] G. Awad, A. Butt, K. Curtis, Y. Lee, J. Fiscus, A. Godil, A. Delgado, J. Zhang, E. Godard, L. Diduch, A. Smeaton, Y. Graham, W. Kraaij, and G. Quenot. TRECVID 2019: An evaluation campaign to benchmark video activity detection, video captioning and matching, and video search retrieval. In *TRECVID Workshop*, 2019.

[3] G. Awad, A. A. Butt, K. Curtis, J. Fiscus, A. Godil, Y. Lee, A. Delgado, J. Zhang, E. Godard, B. Chocot, L. Diduch, J. Liu, Y. Graham, G. J. F. Jones, , and G. Quénot. Evaluating multiple video understanding and retrieval tasks at TRECVID 2021. In *TRECVID Workshop*, 2021.

[4] G. Awad, A. A. Butt, K. Curtis, J. G. Fiscus, A. Godil, Y. Lee, A. Delgado, J. Zhang, E. Godard, B. Chocot, L. L. Diduch, J. Liu, A. F. Smeaton, Y. Graham, G. J. F. Jones, W. Kraaij, and G. Quénot. TRECVID 2020: A comprehensive campaign for evaluating video retrieval tasks across multiple application domains. In *TRECVID Workshop*, 2020.

[5] G. Awad, J. Fiscus, D. Joy, M. Michel, A. Smeaton, W. Kraaij, G. Quénot, M. Eskevich, R. Aly, R. Ordelman, G. Jones, B. Huet, and M. Larson. Trecvid 2016: Evaluating video search, video event detection, localization, and hyperlinking. In *TRECVID Workshop*, 2016.

[6] G. Bertasius, H. Wang, and L. Torresani. Is space-time attention all you need for video understanding? In *ICML*, 2021.

[7] A. Chen, F. Hu, Z. Wang, F. Zhou, and X. Li. What matters for ad-hoc video search? A large-scale evaluation on TRECVID. In *ICCV Workshops*, 2021.

[8] C. Dong, X. Chen, A. Chen, F. Hu, Z. Wang, and X. Li. Multi-level visual representation with semantic-reinforced learning for video captioning. In *ACMMM*, 2021.

[9] J. Dong, X. Li, W. Lan, Y. Huo, and C. G. M. Snoek. Early embedding and late reranking for video captioning. In *ACMMM*, 2016.

[10] J. Dong, X. Li, and C. G. M. Snoek. Predicting visual features from text for image and video caption retrieval. *IEEE Transactions on Multimedia*, 2018.

[11] J. Dong, X. Li, C. Xu, X. Yang, G. Yang, X. Wang, and M. Wang. Dual encoding for video retrieval by text. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[12] J. Dong, Y. Wang, X. Chen, X. Qu, X. Li, Y. He, and X. Wang. Reading-strategy inspired visual representation learning for text-to-video retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022.

[13] C. Feichtenhofer. X3D: Expanding architectures for efficient video recognition. In *CVPR*, 2020.

[14] D. Ghadiyaram, D. Tran, and M. Feiszli. Large-scale weakly-supervised pre-training for video action recognition. In *CVPR*, 2019.

[15] F. Hu, A. Chen, Z. Wang, F. Zhou, and X. Li. Lightweight attentional feature fusion for video retrieval by text. *CoRR*, abs/2112.01832, 2021.

[16] X. Li, J. Dong, C. Xu, J. Cao, X. Wang, and G. Yang. Renmin University of China and Zhejiang Gongshang University at TRECVID 2018: Deep cross-modal embeddings for video-text retrieval. In *TRECVID Workshop*, 2018.

[17] X. Li, S. Liao, W. Lan, X. Du, and G. Yang. Zero-shot image tagging by hierarchical semantic embedding. In *SIGIR*, 2015.

[18] X. Li, C. Xu, G. Yang, Z. Chen, and J. Dong. W2VV++: Fully deep learning for ad-hoc video search. In *ACMMM*, 2019.

[19] X. Li, J. Ye, C. Xu, S. Yun, L. Zhang, X. Wang, R. Qian, and J. Dong. Renmin University of china and Zhejiang Gongshang University at TRECVID 2019: Learn to search and describe videos. In *TRECVID Workshop*, 2019.

[20] X. Li, F. Zhou, and A. Chen. Renmin University of China at TRECVID 2020: Sentence encoder assembly for ad-hoc video search. In *TRECVID Workshop*, 2020.

[21] X. Li, F. Zhou, C. Xu, J. Ji, and G. Yang. SEA: Sentence encoder assembly for video retrieval by textual queries. *IEEE Transactions on Multimedia*, 23:4351–4362, 2021.

[22] Y. Li, Y. Song, L. Cao, J. Tetreault, L. Goldberg, A. Jaimes, and J. Luo. TGIF: A new dataset and benchmark on animated GIF description. In *CVPR*, 2016.

[23] T. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014.

[24] D. Mahajan, R. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, Y. Bharambe, and L. Maaten. Exploring the limits of weakly supervised pretraining. In *ECCV*, 2018.

[25] P. Mettes, D. C. Koelma, and C. G. M. Snoek. Shuffled ImageNet banks for video event detection and search. *ACM Transactions on Multimedia Computing Communications and Applications*, 16(2), 2020.

[26] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. In *ICLR*, 2013.

[27] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[28] N. Reimers and I. Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP*, 2019.

[29] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel. Self-critical sequence training for image captioning. In *CVPR*, 2017.

[30] P. Sharma, N. Ding, S. Goodman, and R. Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018.

[31] D. Tran, L. D. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 2015.

[32] R. Vedantam, C. L. Zitnick, and D. Parikh. CIDER: Consensus-based image description evaluation. In *CVPR*, 2015.

[33] W. Wang, X. Lu, J. Shen, D. Crandall, and L. Shao. Zero-shot video object segmentation via attentive graph neural networks. In *ICCV*, 2019.

[34] X. Wang, J. Wu, J. Chen, L. Li, Y. F. Wang, and W. Y. Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *ICCV*, 2019.

[35] J. Xu, T. Mei, T. Yao, and Y. Rui. MSR-VTT: A large video description dataset for bridging video and language. In *CVPR*, 2016.

[36] X. Yang, K. Tang, H. Zhang, and H. Cai. Auto-encoding scene graphs for image captioning. In *CVPR*, 2019.

[37] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei. Boosting image captioning with attributes. In *ICCV*, 2017.