VIREO @ TRECVid 2021 Ad-hoc Video Search

Jiaxin Wu[†], Zhijian Hou[†], Zhixin Ma[⋆],Chong-Wah Ngo[⋆]

[†]*Department of Computer Science, City University of Hong Kong*
[⋆]*School of Computing and Information Systems, Singapore Management University*
{jiaxin.wu, zjhou3-c}@my.cityu.edu.hk,
zxma.2020@phdcs.smu.edu.sg, cwngo@smu.edu.sg

Abstract

In this paper, we summarize our submitted runs and results for Ad-hoc Video Search (AVS) task at TRECVid 2021 [1].

Ad-hoc Video Search (AVS): We applied three video search systems for AVS: two extended dual-task models trained with a novel unlikelihood loss or with additional phrase concepts, and an enhanced dual encoding model with advanced video features. The unlikelihood model addresses contrary relationships between concepts to prevent embedding features from being interpreted with opposite concepts simultaneously, such as "outdoors" and "indoors". The phrase model extends the concept bank of dual-task model [2] with phrases such that the visual embedding can be interpreted with both words and phrases. The enhanced dual encoding model [3] appends video network input with two advanced deep features extracted from SlowFast [4] and Swin-Transformer [5]. We submitted four automatic runs and four manual runs for both the main task and progress subtask. Besides, we also submitted one novelty run for the main task. We briefly summarize our runs as follows:

- *F_D_C_D_VIREO.21_1*: This automatic run attains the mean xinfAP= 0.317 on the main task and xinfAP= 0.288 on the progress subtask. This run is based on the phrase model, whose concept bank is comprised of words and phrases.
- *F_D_C_D_VIREO.21_2*: This automatic run attains the mean xinfAP= 0.330 on the main task and xinfAP= 0.311 on the progress subtask. This run is based on the unlikelihood model, where contrary relationships of concepts are carefully addressed when generating an interpretation for a visual embedding. This run exhibits the best result on the progress queries over all the submissions.
- *F_D_C_D_VIREO.21_3*: This automatic run obtains the mean xinfAP= 0.336 on the main task and xinfAP= 0.306 on the progress subtask. This run is the ensemble of the phrase model and unlikelihood model, i.e., *F_D_C_D_VIREO.21_1* and *F_D_C_D_VIREO.21_2*.
- *F_D_C_D_VIREO.21_4*: This automatic run attains the mean xinfAP= 0.355 on the main task and xinfAP= 0.305 on the progress subtask. This run is the ensemble of our three search models: the phrase model, the unlikelihood model, and the enhanced dual encoding model. It devotes our best result to main runs.
- *M_D_C_D_VIREO.21_1*: This manual run applies the same system with the same settings presented in the run *F_D_C_D_VIREO.21_1*. The difference is that the original queries are

rephrased with new terms or are reorganized with logical connections via Boolean operators. However, the human intervention degrades the performance of the main task from 0.317 to 0.305 and the performance of the progress subtask from 0.288 to 0.271.

- *M_D_C_D_VIREO.21_2*: This manual run is based on the same system with the same settings presented in the run *F_D_C_D_VIREO.21_2* with manually formulated queries. Similarly, the performance also degrades from 0.330 to 0.301 in the main task and drops from 0.311 to 0.275 in the progress subtask.

- *M_D_C_D_VIREO.21_3*: This manual run is the ensemble of the previous two manual runs. In this run, the performances are degraded by 0.023 xinfAP in the main task and 0.029 in the progress subtask over its corresponding automatic run (i.e., F_D_C_D_VIREO.21_3).

- *M_D_C_D_VIREO.21_4*: This manual run uses the same system with the same settings presented in the run *F_D_C_D_VIREO.21_4* but with manually formulated queries. The performance is maintained as 0.355 in the main task but drops from 0.305 to 0.299 in the progress subtask.

- *M_D_N_D_VIREO.21_5*: Different from the previous runs relying on the fused result of concept-based and concept-free searches, this novelty run only relies on concept searches. It is the ensemble of the concept searches from the phrase model and the unlikelihood model. As a result, this run attains mean xinfAP= 0.297 for the main task along with new manually formulated queries.

1 Ad-hoc Video Search (AVS)

Our previous effort on interpreting visual embedding into semantic meanings has enabled the hybrid search of using both concepts and embeddings for ad-hoc video search (AVS) task [2, 6]. However, there are still some issues overlooked in our previous dual-task model [2]. Firstly, the model can only interpret embeddings into word-level. For example, it interprets the embedding of a video about demonstrating "sign language" as two words, "sign" and "language". Thus, we develop an enhanced model (i.e., the phrase model) by adding phrases derived from video captions to the concept bank. Secondly, the previous model [2] treated each concept independently, which resulted in inconsistent interpretation such as decoding a contrary concept pair "indoors-outdoors" simultaneously for a visual embedding. We address this by proposing a novel unlikelihood loss to encourage coherently decoding concepts by eliminating inconsistent interpretation (namely, unlikelihood model). Thirdly, the video network input (ResNet [7] and ResNeXt [8]) is enhanced with SlowFast [4] and Swin-Transformer [5] features.

1.1 Model descriptions

The core of the phrase model and the unlikelihood model is the dual-task model [2], which not only learns the cross-modal embeddings but also decodes concepts from visual embeddings.

1.1.1 The phrase model

The phrase model enhances the dual-task model with phrase concepts. The new concept bank of the phrase model is comprised of words and phrases. Both of them are derived from the training video captions [9, 10, 11]. We generate a parse tree for each video caption using a constituency parser [12], and store its words and phrases. The process is repeatedly run on all video captions, and we keep the

Figure 1: The word cloud of the new concept bank. The font size indicates the appearance frequency of a concept in the training set.

word/phrase that appears more than 20 times in the concept bank. Lemmatization is involved in the process, and stop words are removed from the vocabulary. As a result, the new concept bank contains 14,528 concepts, including 9,465 phrases and 5,063 words. Figure 1 shows a word cloud of the new concept bank where the font size indicates how frequently a concept appears in the training set. As observed, individual words such as "man" and "woman" appear in most videos, and the phrases such as "young man" and "on stage" show up frequently in videos.

The dual-task model [2] is trained with the new concept bank, and the visual embedding is interpreted with both words and phrases.

1.1.2 The unlikelihood model

The unlikelihood model considers concept relationships, aiming to generate consistent interpretation. Specifically, we define a pair of concepts as globally exclusive if their co-appearance contradicts common sense, such as daytime-nighttime and indoors-outdoors. We also define a pair of concepts that are locally exclusive such as "man-woman", "sit-stand". For example, in a query that finds shots of a short-hair woman, videos with only a short-hair man should be excluded. In interpreting a visual embedding, globally exclusive concepts should not be assigned with both high probabilities. The locally exclusive concepts are allowed to be assigned with both high probabilities only if they are both mentioned in the video captions or metadata.

Inspired by the unlikelihood training in NLP [13, 14], we propose a novel unlikelihood loss in video domain to address the issue of exclusive relationships between concepts in visual embedding interpretation. Given a predicted probability $\hat{p} = [\hat{p}_1, \hat{p}_2, ..., \hat{p}_i, ..., \hat{p}_n] \in \mathbb{R}^{n+}$ for a visual embedding and its ground truth $p = [p_1, p_2, ..., p_i, ..., p_n] \in \{1, 0\}^n$, the unlikelihood loss is computed as:

$$Loss_{UL}(\hat{p}, p) = \frac{1}{\sum_i^n p_i} \sum_i^n -p_i \sum_{t \in T_i} log(1 - \hat{p}_t) * (1 - p_t) \quad (1)$$

where n is the size of concept bank and T_i is a set of exclusive concepts for the concept i. We employ the likelihood loss originally proposed in [2] and the proposed unlikelihood loss together to generate a

consistent interpretation for visual embeddings.

1.1.3 The enhanced dual encoding model

Original dual encoding model [3] learns three levels of information (mean pooling, biGRU, and biGRU-CNN), concatenates all information into a single representation, and finally maps the representation into the joint embedding space for both video and textual query input. The core idea of the enhanced dual encoding model is the adoption of two enhanced video features. Note that the above phrase and unlikelihood models also adopt the enhanced video features.

First, we extract addition motion feature SlowFast [4] pre-trained on Kinetics [15]. We notice that the previous appearance features are not sufficient to capture some queries which focus more on the motion-level information. For example, the query-653 *Find shots of group of people clapping* emphasizes the clapping motion. It is not enough to use key-frame appearance features to determine whether a group of people are clapping or not, since clapping is a continuous hand-moving state.

Furthermore, we also enhance the appearance feature. Inspired by the impressive performance of Swin-Transformer [5] for various computer vision tasks, we further extract the Swin-Transformer feature. It is pre-trained on ImageNet-21K and finetuned on ImageNet-1K and shows competitive performance on ImageNet leaderboard.

To this end, we extract four features (ResNet, ResNeXt, Swin-Transformer, SlowFast) to capture diverse video information and use them as video network input.

2 Results analysis

In this year's AVS benchmarking, the evaluation is conducted on the V3C1 dataset [16] and two groups of queries. The first group contains twenty new main queries released this year. The other group includes ten progress queries released in 2019. The use of progress queries is to evaluate the progress of the submitted systems in three continuous years, i.e., 2019-2021.

Figure 2 shows the mean extended inferred average precision (xinfAP) of our submissions on 20 new queries this year. Our run, *F_D_C_D_VIREO.21_4*, achieves the top-1 performance out of 30 submissions on the main queries. Specifically, we obtain the highest xinfAP scores on 8 out of 20 queries. Some of them benefit from the interpretability of our model. For example, for the query-679 *Find shots of a ladder with less than 6 steps*, the average performance of the submitted system is 0.04 xinfAP. However, as our models are able to interpret the embeddings as ["person", "climb", "ladder", "stairs", "steps"], the average xinfAP of our runs is 0.108. Besides, some of the queries are improved by having phrases. For instance, for the query-676 *Find shots of a white dog*, as "white dog" is a phrase in the vocabulary,

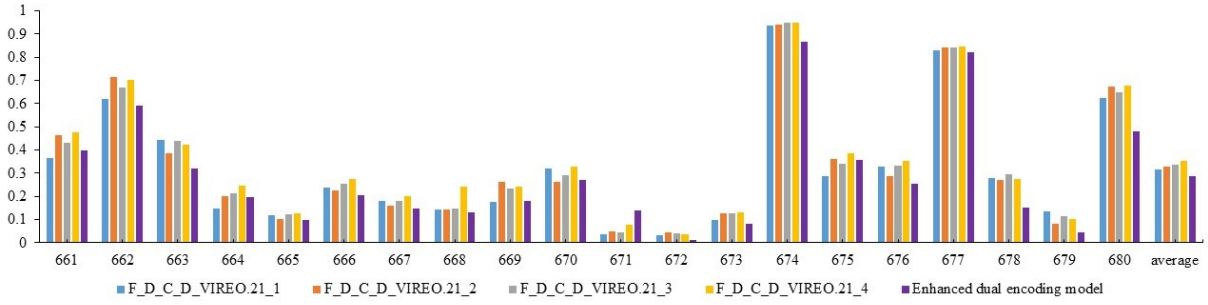

Figure 2: Performance comparison of our submissions in fully-automatic runs on main queries.

Table 1: The performance comparison of our models on tv21 query set

Id	Model	Concept search	Embedding search	Fusion search
Baselines:				
m1	dual-task model [2]	0.167	0.167	0.193
m2	dual-task model + enhanced features	0.269	0.278	0.305
Proposed models:				
m3	phrase model (RUN1)	0.216	0.301	0.317
m4	unlikelihood model (RUN2)	0.270	0.290	0.330
m5	enhanced dual encoding model	—	0.287	—
Ensemble models:				
m6	m3+m4 (RUN3)	—	—	0.336
m7	m3+m4+m5 (RUN4)	—	—	0.355

the concept search of our phrase model obtains xinfAP=0.425, which exceeds other submissions by about 0.1 xinfAP. The unlikelihood model also brings improvements to some queries. For example, for the query-673 *Find shots of a man pointing with his finger*, our unlikelihood model prunes video segments of woman out of the ranking list and doubles the average performance of submitted run on this query.

Table 1 further compares the performances of our models on the tv21 query set. As observed, with appending visual features (i.e., Slowfast and Swin-Transformer) and additional training set (i.e., VATEX [11]), the performance significantly boosts beyond the original dual-task model [2] on all search modes. Note that all our proposed models are trained with the enhanced features and the additional training dataset. In the comparison of the concept search, the phrase model is surprisingly worse than the model m2, which only has single words as concepts on tv21. It is because most important phrases are not well trained. For example, for the query-662 *Find shots of a woman wearing sleeveless top*, using the combination of the single words "sleeveless" and "tops" obtains xinfAP=0.274 while using the phrase "sleeveless tops" obtains xinfAP=0.007. The reason is "sleeveless tops" only has 28 training cases which are not enough to train a good classifier. Similar situations happen in the phrases "hang glider", "wear blue jacket", and "barber chair". However, having phrases as interpretation still boosted the embedding search and fusion search performance as both single words and phrases are used for embeddings. Compared with the concept-based search of the original dual-task model, the unlikelihood model works particularly well on those queries which involve exclusive concepts, e.g., query-662 *Find shots of a woman wearing sleeveless top* and query-676 *Find shots of a white dog*. The unlikelihood training has removed videos of man for query-662 and videos of black dog for query-676, and the performances are risen by about 100% and 10%, respectively. In terms of concept-free enhanced dual encoding model (m5), it works better on query-671 *Find shots of a man behind a pub bar or club bar*, dual-task models can not learn the knowledge of inter-object spatial location relationship "behind", and find many incorrect videos where a man is in front of club bar. In contrast, the concept-free model implicitly learns the spatial location relationship via embedding space. Furthermore, the predictions of concept-free model (m5) and the ensemble of two dual-task variant models (m6) are complementary on 14 out of 20 queries. The query-668 *Find shots of a person wearing an apron indoors* enjoys the most benefit of model ensemble. The m5 or m6 can only achieve xinfAP around 0.14 separately, while the three-model ensemble (m7) achieves xinfAP=0.241.

Our models also work well on progress queries with three runs being ranked within the top-3 performing submissions. Specifically, our unlikelihood model ranks at the first position, and our two ensemble

(a) Automatic run (b) Manual run

Figure 3: Performance comparison of the automatic run and the manual run on the query-607 *Find shots of two people kissing who are not bride and groom.*

models (i.e., m6 and m7) followed closely behind. Ensemble approach, *F_D_C_D_VIREO.21_4*, does not make improvements because the components (i.e., the phrase model and the enhanced dual encoding model) did not work well on progress queries. For example, for the query-601 *Find shots of a person jumping with a motorcycle*, these two models obtain 0.182 and 0.125, respectively, while the unlikelihood model obtains 0.346. Besides, significant improvements can be seen by the latest systems. Our systems in 2021 (xinfAP=0.303) outperform our systems in 2020 (xinfAP=0.199) and 2019 (xinfAP=0.146) by a large margin. Taking query-608 *Find shots of two people talking to each other inside a moving car* as an example, our systems in 2019 are all concept-based, and they failed miserably in this query which obtained an average performance of 0.01, because the concept-based systems failed in detecting "two people", "talking", and "moving car". In 2020, the number was improved to 0.05 by our embedding model as more shots of two people are found while they are not talking to each other. This year, by having the concept phrase "two people" and additional motion features, we obtained a more satisfactory score with xinfAP=0.139.

Besides the automatic runs, we also submitted manual runs. Our manual runs are not able to exceed the automatic runs due to inconsistent performance across queries. Table 2 lists the original queries and the modified queries and their performances of using the unlikelihood model on main queries. For example, when reformulating the query-671 *Find shots of a man behind a pub bar or club bar* to *Find shots of bartender man making alcohol in a bar*, the performance is up by seven times. However, when the query is modified from *Find shots a person painting on a canvas* to *Find shots a person painting on a canvas stock*, the score drops from 0.262 to 0.107 because the model fails in finding canvas stock. We also use Boolean operation for some queries, such as using NOT in the query-607 *Find shots of two people kissing who are not bride and groom.* Specifically, the retrieved scores of two sub-queries *Find shots of two people kissing* and *Find shots of bride and groom* are generated, and the final ranking score is obtained by subtracting the latter from the former. Their retrieved top-40 rank lists are visualized in Figure 3. The red border marks the wrong segments, while the green marks the ground truth video segments. The Boolean operator manages to obtain improvement in this query. However, for other modified queries having additional NOT such as query-663, query-666, their performances are decreased. As observed, when pruning out the false positives, true positives also have a risk of being removed as well, and the performance depends on the effectiveness of the pruned classifier. It is also interesting that, for query-677, the performance decrease because video segments of a jewelry ring in the box contaminate the result for the modified query finding shots of two boxers in ring (where "a" is removed from the original query).

Furthermore, our novelty run also obtains a competitive performance with our other runs. The novelty

Table 2: Performance comparison of the original queries and the modified queries using the unlikelihood fusion search, i.e., *F_D_C_D_VIREO.21_2* versus *M_D_C_D_VIREO.21_2*.

original query	xinfAP	modified query	xinfAP
661 a hang glider floating in the sky on a sunny day	0.464	661 wing hang gliders with sun	0.293
662 a woman wearing sleeveless top	0.713	662 a woman wearing sleeveless top not wearing dress	0.265
663 a person with a tattoo on their arm	0.386	663 a person with a tattoo on their arm not on their back	0.323
664 city street where ground is covered by snow	0.201	664 snow falling on city street	0.191
665 an adult person wearing a backpack and walking on a sidewalk	0.103	665 a person walking on sidewalk with backpack	0.112
666 a man wearing a blue jacket	0.227	666 a man wearing a blue jacket not blue shirt	0.193
667 a person looking at themselves in a mirror	0.158	667 a person in front of a mirror	0.179
668 a person wearing an apron indoors	0.142	668 a person wearing an apron in the kitchen	0.111
669 a woman holding a book	0.262	669 a woman holding a book in the library	0.172
670 a person painting on a canvas	0.262	670 a person painting on a canvas stock	0.107
671 a man behind a pub bar or club bar	0.047	671 bartender man making alcohol in a bar	0.370
672 a person wearing a cap backwards	0.045	672 rapper wearing a backward cap	0.162
673 a man pointing with his finger	0.127	673 a man pointing out with his finger	0.119
674 a parachutist descending towards a field on the ground in the daytime	0.942	674 a parachute landing field grass	0.969
675 two or more ducks swimming in a pond	0.362	675 many ducks swimming in a pond water	0.260
676 a white dog	0.288	676 a white dog not a white cat	0.259
677 two boxers in a ring	0.840	677 two boxers in ring	0.798
678 a man sitting on a barber chair in a shop	0.272	678 a man has haircut barber shop	0.403
679 a ladder with less than 6 steps	0.081	679 a short step ladder	0.086
680 a bow tie	0.672	680 bowtie	0.648
average	0.330		0.301

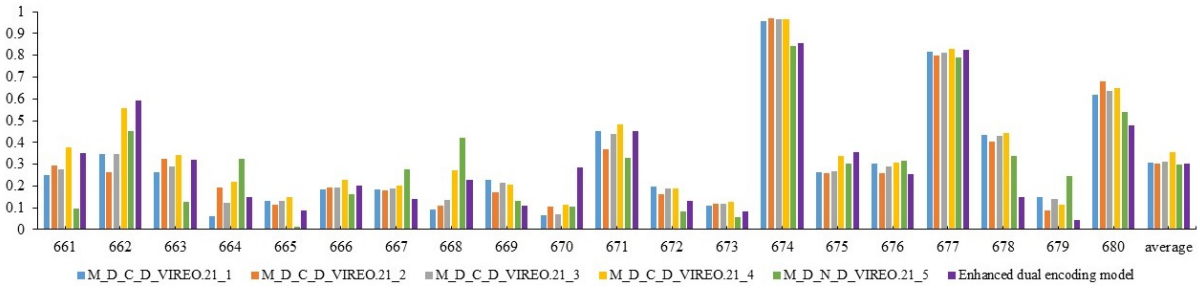

Figure 4: Performance comparison of our manual runs on the main queries.

run, *M_D_C_D_VIREO.21_5*, is the ensemble of two concept searches from the unlikelihood model and the phrase model. In order to find unique video segments, the input queries used in the novelty run have some differences from those listed in Table 2. The differences are presented in Table 3. Figure 4 compares our novelty run with other submitted runs and the enhanced dual encoding model on the main queries with manual setting. On average, our novelty run obtains a mean xinfAP of 0.297, which is competitive with the embedding search (i.e., enhanced dual encoding model) and fusion search (i.e., Run1-Run4). Specifically, our novelty run manages to obtain the best result on 4 out of 20 queries, i.e., query-664, query-667, query-668, and query-679. It is interesting to find that all of these four queries are simplified as finding shots of objects. For example, for the query-668 *Find shots of a person wearing an apron indoors*, after simplifying this query as finding shots of an apron, the novelty run obtains a xinfAP of 0.422, which is the highest performance among all submissions this year. A similar situation happens in the query-679 *Find shots of a ladder with less than 6 steps*. The novelty run manages to double the second-best performance this year by only using "ladder" in retrieval. To verify whether the improvements come from the modified queries or concept-based search, we test the embedding searches of the unlikelihood and the phrase models with these four modified queries input to the novelty run. The results show that, except query-667, the embedding searches are not competitive with concept search on finding these objects by obtaining half of the xinfAP scores on these queries.

Table 3: The different manual queries input in novelty run.

Original query	Modified query
664 city street where ground is covered by snow	664 city street with snow
665 an adult person wearing a backpack and walking on a sidewalk	665 person hiking on sidewalk with backpack
667 a person looking at themselves in a mirror	667 person and mirror
668 a person wearing an apron indoors	668 an apron
671 a man behind a pub bar or club bar	671 bartender
672 a person wearing a cap backwards	672 a backward cap
677 two boxers in a ring	677 box ring
679 a ladder with less than 6 steps	679 ladder
680 a bow tie	680 a bow tie not on wedding

3 Conclusion

Our study this year aims to address the limitations of dual-task model towards learning interpretable embedding. First, the proposed unlikelihood training is verified to be effective by downgrading the rank of videos with concepts which contradict to the constraints specified in a query. Second, being able to interpret an embedding with phrases also makes the learnt embedding features more robust to search. However, using the interpreted phrases instead of words for concept-based search plays the risk that the performance can drop substantially if the number of training examples available for a phrase is very few. Finally, the use of more advanced features (SlowFast motion-based and Swin-Transformer image-based features) significantly boosts our performance this year. Overall, our results on the progress queries are noticeably better than the results in two previous years.

For manual run and novelty run, the problem where the results are sensitive to how a query is expressed remains. The problem is multi-faceted and depends on the training data and the video dataset. For example, making a query more specific with manually adding terms (e.g., adding *kitchen* to *a person wearing an apron* will downgrade the search performance. Abbreviating the query to *an apron*, instead, "frustratingly" improves the performance. We believe that these issues will be better addressed in interactive search and our result of manual run is somewhat "trying luck" or arbitrary.

4 Acknowledgment

This research was supported by the Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 1 grant.

References

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

[2] J. Wu and C.-W. Ngo, "Interpretable embedding for ad-hoc video search," in *Proceedings of the ACM Conference on Multimedia*, 2020.

[3] J. Dong, X. Li, C. Xu, S. Ji, Y. He, G. Yang, and X. Wang, "Dual encoding for zero-example video retrieval," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[4] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proceedings of the IEEE/CVF ICCV*, 2019, pp. 6202–6211.

[5] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

[6] J. Wu, P. A. Nguyen, Z. Ma, and C.-W. Ngo, "Sql-like interpretable interactive video search," in *MultiMedia Modeling*, 2021, pp. 391–397.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, 2015.

[8] S. Xie, R. Girshick, P. Dollar, Z. W. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[9] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[10] J. Xu, T. Mei, T. Yao, and Y. Rui, "Msr-vtt: A large video description dataset for bridging video and language," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2016.

[11] 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 *The IEEE International Conference on Computer Vision*, 2019.

[12] M. Gardner, J. Grus, M. Neumann, O. Tafjord, P. Dasigi, N. F. Liu, M. Peters, M. Schmitz, and L. S. Zettlemoyer, "Allennlp: A deep semantic natural language processing platform," 2017.

[13] M. Li, S. Roller, I. Kulikov, S. Welleck, Y.-L. Boureau, K. Cho, and J. Weston, "Don't say that! making inconsistent dialogue unlikely with unlikelihood training," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 4715–4728.

[14] S. Welleck, I. Kulikov, S. Roller, E. Dinan, K. Cho, and J. Weston, "Neural text generation with unlikelihood training," in *ICLR*, 2019.

[15] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, "The kinetics human action video dataset," *CoRR*, vol. abs/1705.06950, 2017. [Online]. Available: http://arxiv.org/abs/1705.06950

[16] F. Berns, L. Rossetto, K. Schoeffmann, C. Beecks, and G. Awad, "V3c1 dataset: An evaluation of content characteristics," in *Proceedings of the International Conference on Multimedia Retrieval*, 2019, pp. 334–338.