RUC_AIM3 at TRECVID 2023: Video to Text Description

Kaiwen Wei, Zihao Yue, Liang Zhang, Qin Jin*
School of Information, Renmin University of China
{kaiwenwei, yzihao, zhangliang00, qjin}@ruc.edu.cn

Abstract

This report presents our solution for the Video to Text Description (VTT) task of TRECVID 2023. Based on our baseline VTT model in TRECVID 2022, we further improve the captioning performance by leveraging a more advanced video-text pretraining model, augmenting the training with more high-quality video-text data, and applying a re-ranking strategy for top candidate caption selection. Our submissions from our improved VTT model rank the 1^{st} in TRECVID VTT 2023 on evaluation metrics including CIDErD, CIDEr, METEOR and STS in the main task, achieving the best CIDEr of 39.4.

1 Introduction

Video to Text Description (VTT) is a challenging vision-language task, which aims to automatically generate natural language descriptions given short videos (Awad et al., 2023). The mainstream solutions for the VTT task usually rely on image-text pre-training models (Zhang et al., 2021b; Radford et al., 2021; Li et al., 2022),. For example, Yue et al. (2022) fine-tune an image-text pre-training model BLIP (Li et al., 2022) on video data, achieving promising results on the VTT task. This demonstrates the powerful visual understanding and textual generation capabilities of image-text models can be effectively transferred to video tasks. However, building video description systems from image-text models leads to limitations in temporal modeling.

Hence, we consider using a video-text pretraining model with a temporal understanding module for better video captioning. Specifically, we apply mPLUG-2 (Xu et al., 2023) as our basic captioning model because it achieves SOTA video captioning performance on MSRVTT (Xu et al., 2016) through large-scale video-text pre-training. To better fine-tune the model, we improve both the quality and quantity of the training dataset through pseudo-labeling and back translation. We generate multiple candidate descriptions for each video and employ a re-ranking method to select the best one. With the above-mentioned components, our system ranks the 1^{st} place in TRECVID VTT 2023.

* Corresponding author

2 Related Work

To generate video descriptions, early works rely on off-the-shelf feature extractors to get video representations. For example, Venugopalan et al. (2015) generated video descriptions through LSTM (Hochreiter and Schmidhuber, 1997) by accepting video features from CNN (LeCun et al., 1998). Zhang et al. (2021a) ensembles several types of features within transformer architecture (Vaswani et al., 2017). These methods suffer from weak visual understanding since their video representations are kept fixed and limited by the feature extractors.

Later, image-text pre-training models (Zhang et al., 2021b; Radford et al., 2021; Li et al., 2022) show strong visual understanding abilities through learning from large-scale paired image-text data on the web (Sharma et al., 2018; Schuhmann et al., 2022). Many efforts try to transfer the capability of image-text models to video tasks (Liu et al., 2022; Yan et al., 2023; Liu et al., 2023). Specifically, He et al. (2023) proposes a video adapter module to empower CLIP (Radford et al., 2021) with temporal modeling. Yue et al. (2022) build a video captioning system on the basis of BLIP (Li et al., 2022). Though significant improvements are made, these models have limited ability to capture motion in videos (Li et al., 2022), since their backbone captioning models are pre-trained with static images.

More recently, with the rapid development of VLP, many works (Wang et al., 2023; Xu et al., 2023) introduce video-text data during pre-training. This enables their models to have stronger temporal modeling abilities and perform better in video

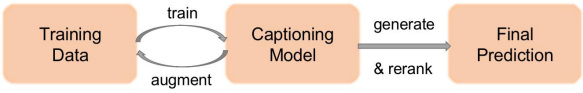

Figure 1: Our overall framework.

tasks. For example, Xu et al. (2023) incorporate both image-text and video-text data for pre-training and propose a modularized model mPLUG-2. It achieves SOTA in video description generation on the MSRVTT dataset (Xu et al., 2016). We thus choose mPLUG-2 as the backbone captioning model in our system.

3 Method

As illustrated in Fig. 1, our VTT system mainly consists of three components: the captioning model, the data augmentation module, and the candidate re-ranking module. The captioning model generates descriptions for given videos. The data augmentation module creates and filters pseudo video descriptions to re-train the model. The candidate re-ranking module assigns a quality score for each candidate description generated by our system. During inference, our system first generates multiple candidate captions by the trained captioning model. With our re-ranking module, we then evaluate the quality of each candidate caption and select the best one as the final prediction.

3.1 Captioning Model

Temporal information is essential for generating video descriptions. Our winning system in TRECVID 2022 (Yue et al., 2022) chooses the image-text pre-training model BLIP (Li et al., 2022) as the captioning model. It gains little temporal information understanding abilities by small-scale fine-tuning. In contrast, video-text pre-training models could obtain temporal modeling ability from large-scale pre-training. Thus, they could capture motion information in the video better and generate more accurate descriptions. We therefore utilize the state-of-the-art video-text pre-training model mPLUG-2 (Xu et al., 2023) as our backbone model to replace BLIP.

Given a video, mPLUG-2 first samples keyframes, and extracts visual features with a dual-vision encoder and a universal layer. It then feeds the visual features into a decoder to generate video descriptions. Compared to image-text models, mPLUG-2 has the potential to capture temporal information for its spatial-temporal modeling module in the dual-vision encoder.

If not otherwise specified, we fine-tune mPLUG-2 in two stages to make it fit the VTT dataset better. We first optimize the model with cross-entropy loss in the first stage and then adopt reinforcement learning with SCST (Rennie et al., 2017) in the second stage.

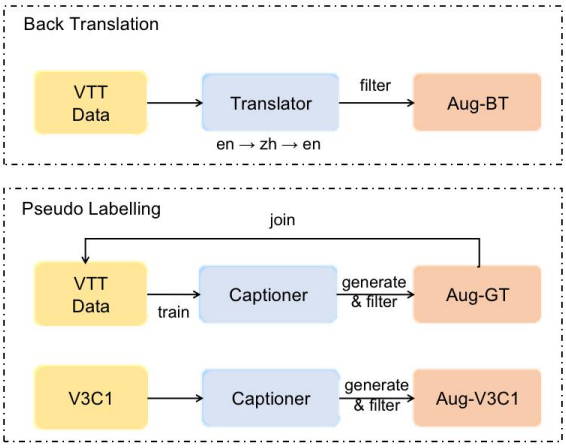

Figure 2: Our data augmentation pipeline.

3.2 Data Augmentation

To further improve our model, we apply data augmentation to obtain additional high-quality training data. As shown in Fig. 2, our data augmentation pipeline comprises back translation and pseudo labeling.

Back Translation (Sennrich et al., 2015) is applied to increase the diversity of the existing video descriptions. Specifically, we translate ground truth captions to Chinese with Baidu Translation (He, 2015). Then, we translate them back to English. We filter the back-translated captions with CIDEr scores to ensure their quality.

Pseudo Labeling leverages the captioning model to create pseudo descriptions from videos. These pseudo descriptions are then filtered and added to the training data to improve the captioning model. We can cycle through the above procedure since the improved model can continue to be applied for pseudo-labeling. The details of our pseudo-labelled data are shown in Section 4.2.

3.3 Re-ranking

Re-ranking aims to select the best caption from multiple candidate captions. In our previous solu-

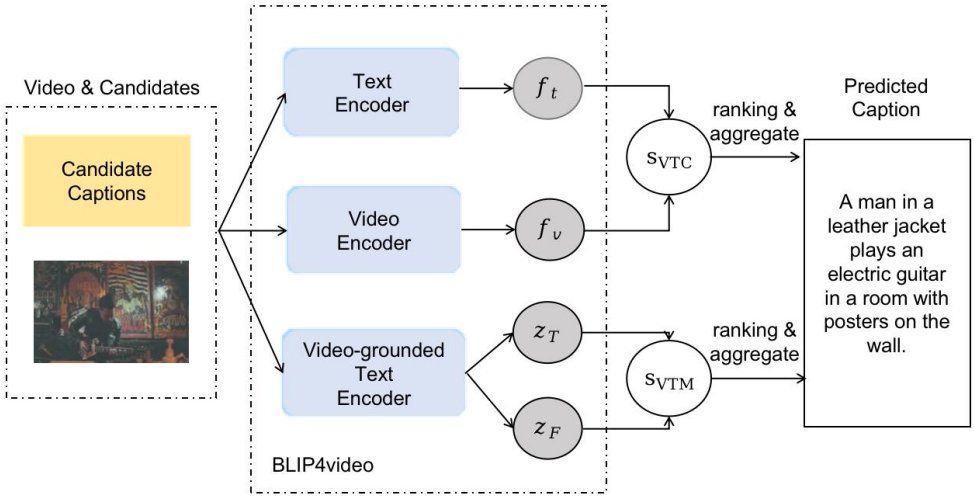

Figure 3: Our re-ranking strategy.

tion (Yue et al., 2022), we adopt a visual-grounded text encoder (Li et al., 2022) to calculate Video-Text Matching (VTM) scores. The encoder outputs z_T and z_F, representing confidence scores for whether the video and text match (T) or not (F). Then, we obtain the probability of video matching text as the VTM score by softmax:

$$s_{\text{VTM}} = \frac{e^{z_T}}{e^{z_T} + e^{z_F}} \tag{1}$$

With token-level cross-attention between video and text, VTM scores measure fine-grained video-text alignments. However, we believe that additional consideration of the overall semantic similarity of video and text can be beneficial. Thus, in addition to VTM scores, we propose to take into account the Video-Text Contrastive (VTC) score as well. VTC scores depict the cosine similarity of video and text features extracted by a contrastive learning-based video-text retrieval model. VTC scores can be calculated as follows:

$$s_{\text{VTC}} = \frac{f_v \cdot f_t}{|f_v||f_t|} \tag{2}$$

Where f_v is the video feature and f_t is the text feature. Given a candidate, we sum the rankings of its VTM and VTC score, and a smaller ranking sum implies a better caption quality. We adopt BLIP4video (Yue et al., 2022) fine-tuned on VTT16-21 with video-text contrastive and video-text matching tasks as the scoring model for both VTM and VTC scores. We select a caption with the smallest ranking sum of VTM and VTC as the output of our captioning system.

For candidate generation, we consider a pipeline involving i distinct models. Each model generates $j \times k$ candidates for each video, where we sample frames using TSN sampling j times to get different frame inputs, and the model generates k beams for each input. The overall framework of our re-ranking strategy is shown in Fig. 3.

4 Experiment

4.1 Captioning Model

To verify the performance of our chosen captioning model mPLUG-2, we perform a shallow evaluation to compare it with last year's captioning model, BLIP4video (Yue et al., 2022) under zero-shot and supervised settings. For the supervised setting, we fine-tune both models on the VTT16-21 data. As shown in Table 1, under both zero-shot and supervised settings, mPLUG-2 significantly outperforms BLIP4video, suggesting mPLUG-2 as a stronger captioning model for the VTT task. We also report its performance after CIDEr optimization, which demonstrates that it can be further improved with SCST (Rennie et al., 2017).

Implementation details. The implementation of our backbone model follows largely the official implementation of mPLUG-2. For video inputs, we extract 16 frames with TSN sampling (Wang et al., 2016). For text outputs, we generate 3 captions via beam search with a beam size of 5, and the minimum generation length is set to 18. During cross-entropy loss fine-tuning, we train the models

Table 1: CIDEr scores of mPLUG-2 and BLIP4video on VTT22. VTT 16-21 are used for fine-tuning.

Approach	Model	CIDEr
Zero-shot	BLIP4video	28.9
	mPLUG-2	**44.8**
Fine-tuned	BLIP4video	50.5
	mPLUG-2	54.4
	mPLUG-2 + SCST	**57.1**

Table 2: Training data of our 4 captioners. CE refers to cross-entropy, and SCST refers to self-critical sequence training.

Model	Training data	
	CE	SCST
Cap-0	VTT16-21	VTT18-21
Cap-1	*Aug-1*	-
Cap-2	*Aug-1*	VTT18-21
Cap-3	*Aug-2*	VTT18-21

for 10 epochs with a batch size of 32. The optimizer is AdamW, and the learning rate for the vision encoder and other modules are $1e{-}7$ and $1e{-}6$, respectively. For SCST, the models are trained for 5 epochs with a batch size of 16 and a learning rate of $5e{-}8$ for all parameters.

4.2 Data Augmentation

Implementation Details. We perform a cyclic data augmentation pipeline as follows: (1) With the officially provided checkpoint of mPLUG-2, which is pre-trained with large-scale data and fine-tuned on MSRVTT (Xu et al., 2016), we further fine-tune it on VTT16-21 by both cross-entropy loss and self-critical sequence training to get a model *Cap-0*. (2) With *Cap-0* as the captioner, we generate 3 captions for each of VTT16-21 videos as pseudo labels, which are added to the original training data (VTT16-21). We name the augmented data set *Aug-1*. We also include the augmentation data provided by Yue et al. (2022). We set a threshold CIDEr $>$ 55 to filter pseudo captions for *Aug-1*. (3) We fine-tune mPLUG-2 on *Aug-1* to get two new captioners, namely *Cap-1* (without SCST) and *Cap-2* (with SCST). (4) We use *Cap-1* and *Cap-2* as captioners to generate a new batch of pseudo labels. At this round, we extend our video source to a subset of the V3C1 containing a randomly selected fifth of the videos[1], in addition to VTT16-21. We also conduct back-translation on the VTT16-21 ground truth captions. These richly sourced augmentation data make up *Aug-2*. For *Aug-2*, we set the threshold for augmentation captions as CIDEr $>$ 80 for VTT videos, and VTM $>$ 60 for V3C1 videos. (5) Finally, we fine-tune mPLUG-2 on *Aug-2* by cross-entropy and self-critical sequence training to get *Cap-3*. More details about the augmentation data are shown in Table 5.

[1]We only consider videos that are 5-15 seconds in length.

Table 3: Details of our augmentation data.

Augment	Data	Description
Aug-1	VTT16-21	VTT data from 2016 to 2021
	Aug-22	Augmentation data from Yue et al. (2022)
	Aug-GT-1	Pseudo labeling for VTT16-21 by *Cap-0*
Aug-2	VTT-22	VTT data 2022
	Aug-BT	Back translation for VTT16-21
	Aug-GT-2	Pseudo labeling for VTT16-21 by *Cap-1*
	Aug-GT-3	Pseudo labeling for VTT16-21 by *Cap-2*
	Aug-V3C1	Pseudo labeling for V3C1 by *Cap-2*

Finally, we obtain 86,078 captions in *Aug-1* and 91,194 captions in *Aug-2*, greatly increasing the training data scale. To verify the effect of different sources of data augmentation, e.g., back-translation, pseudo labels of V3C1, we perform an ablation study of data recipes. As shown in Table 6, back-translated data are more effective than the V3C1 pseudo labels, while pooling all the augmentation data leads to the best performance.

4.3 Re-ranking

Our re-ranking implementation includes 2 models, *Cap-2* and *Cap-3*. Each model generates 5×3 captions per video, where we randomly sample frames 5 times by TSN sampling as input, and generate 3 captions for each input. We ablate the scoring scheme and captioners and demonstrate the results in Table 7. Comparing Row 1 and Row 3, re-ranking across two models performs better than with a single *Cap-3* model. A comparison of Row 2 and Row 3 suggests that combining VTM and VTC scores is more effective than using VTC alone.

4.4 Main Results

The validation performance (VTT 2022) and the official evaluation results on the VTT 2023 test set of our final runs are detailed in Table 4. Despite the narrow variance in performance across the four runs, Run4, which adopts comprehensive augmented data and a complete re-ranking strategy, stands out marginally. Finally, our submission ranks the 1^{st} on evaluation metrics including CIDErD, CIDEr, METEOR and STS in TRECVID

Table 4: Official evaluation results of our submissions.
C: CIDEr, B@4: BLEU@4, M:METEOR, SP:SPICE, ST:STS

Submission	Captioner		Re-ranking		Main Task					Robust Task				
	Cap-2	Cap-3	VTM	VTC	C	B@4	M	SP	ST	C	B@4	M	SP	ST
run1	✓		✓	✓	38.4	9.21	32.81	14.9	47.0	38.9	9.41	33.05	14.8	20.52
run2		✓	✓	✓	**39.4**	9.45	**33.25**	**15.2**	47.3	38.6	9.68	33.04	15.0	20.50
run3	✓	✓		✓	**39.4**	**9.48**	33.19	15.1	47.3	38.4	9.72	33.15	14.9	20.36
run4	✓	✓	✓	✓	**39.4**	**9.48**	33.16	**15.2**	**47.4**	**39.0**	**9.83**	**33.24**	**15.1**	**20.61**

Table 5: Composition and filtering criterion of Aug-1 and Aug-2.

Data	*Aug-1*		*Aug-2*	
	Count	Filter	Count	Filter
VTT16-21	45,820	-	51,820	-
Aug-22	34,660	CIDEr > 55	8,902	CIDEr > 80
Aug-GT	5,598	CIDEr > 55	13,220	CIDEr > 80
Aug-V3C1	-	-	4,392	VTM > 60
Aug-BT	-	-	12,860	CIDEr > 80
Total	86,078	-	91,194	-

Table 6: Performance on VTT22 by using different parts of our augmentation data.

Model	VTT Data	Aug-GT	Aug-BT	Aug-V3C1	CIDEr
Cap0	✓				57.1
Cap2	✓	✓			59.5
Cap2+	✓	✓		✓	59.6
Cap2+	✓	✓	✓		60.0
Cap3	✓	✓	✓	✓	**61.0**

VTT 2023, with the highest CIDEr score of 39.4.

We also submit to the robustness sub-task, which introduces natural corruptions and perturbations to videos, e.g., spatial-temporal corruptions and different types of noise. As shown in Table 4, our models achieve basically the same performance as the main task, indicating they are robust enough to handle these perturbations. We consider the models can benefit from the video input augmentation integrated into the original mPLUG-2, including strategies like random cropping and random frame selection, which likely enhance their robustness to such disturbances.

5 Conclusion

This report presents our solution for the VTT challenge in TRECVID 2023. We adopt a powerful vision-text pre-training model mPLUG-2 as the backbone to generate high-quality video descriptions. To enlarge the training data, we introduce a well-designed data augmentation pipeline with pseudo-labeling and back translation. Lastly, we select the best candidate from multiple generated descriptions with re-ranking strategies. Experiments demonstrate the effectiveness of our designs, and our submissions rank 1st on both the main task and the robust sub-task.

Table 7: Performance on VTT22 by using different re-ranking strategies.

Row	Re-rank		Captioner		CIDEr
	VTM	VTC	Cap-2	Cap-3	
1	✓	✓		✓	62.1
2		✓	✓	✓	62.5
3	✓	✓	✓	✓	**63.1**

6 Discussions

Robustness Subtask. In an optimal benchmark scenario, we would expect the captioning model's performance to decline when processing corrupted videos. Contrary to this expectation, our system demonstrates comparable efficacy on the main task as well as the robustness subtask, suggesting that the robustness subtask does not present a significant challenge. We postulate that the automatically introduced corruptions in the subtask (e.g., noise and compression artifacts) may not significantly impede contemporary AI systems. A likely source of more suitable robustness challenges lies in real-world video recording conditions, such as inadequate lighting and camera shake.

Acknowledgements

This work was partially supported by the National Natural Science Foundation of China (No. 62072462) and the National Key R&D Program of China (No. 2020AAA0108600).

References

George Awad, Keith Curtis, Asad A. Butt, Jonathan Fiscus, Afzal Godil, Yooyoung Lee, Andrew Delgado, Eliot Godard, Lukas Diduch, Deepak Gupta,

Dina Demner Fushman, Yvette Graham, and Georges Quénot. 2023. Trecvid 2023 - a series of evaluation tracks in video understanding. In *Proceedings of TRECVID 2023*. NIST, USA.

Xingjian He, Sihan Chen, Fan Ma, Zhicheng Huang, Xiaojie Jin, Zikang Liu, Dongmei Fu, Yi Yang, Jing Liu, and Jiashi Feng. 2023. Vlab: Enhancing video language pre-training by feature adapting and blending. *arXiv preprint arXiv:2305.13167*.

Zhongjun He. 2015. Baidu translate: Research and products. In *Proceedings of the Fourth Workshop on Hybrid Approaches to Translation (HyTra)*, pages 61–62.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.

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

Yuqi Liu, Luhui Xu, Pengfei Xiong, and Qin Jin. 2023. Token mixing: parameter-efficient transfer learning from image-language to video-language. In *Proc. of AAAI*, volume 37, pages 1781–1789.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *Proc. of CVPR*, pages 7008–7024.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Improving neural machine translation models with monolingual data. *Proc. of ACL*.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. 2015. Sequence to sequence-video to text. In *Proc. of ICCV*, pages 4534–4542.

Alex Jinpeng Wang, Yixiao Ge, Rui Yan, Ge Yuying, Xudong Lin, Guanyu Cai, Jianping Wu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. 2023. All in one: Exploring unified video-language pre-training. *Proc. of CVPR*.

Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. 2016. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer.

Haiyang Xu, Qinghao Ye, Ming Yan, Yaya Shi, Jiabo Ye, Yuanhong Xu, Chenliang Li, Bin Bi, Qi Qian, Wei Wang, Guohai Xu, Ji Zhang, Songfang Huang, Fei Huang, and Jingren Zhou. 2023. mplug-2: A modularized multi-modal foundation model across text, image and video. *Proc. of ICML*.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proc. of CVPR*, pages 5288–5296.

Shen Yan, Tao Zhu, Zirui Wang, Yuan Cao, Mi Zhang, Soham Ghosh, Yonghui Wu, and Jiahui Yu. 2023. Videococa: Video-text modeling with zero-shot transfer from contrastive captioners.

Zihao Yue, Yuqi Liu, Liang Zhang, Linli Yao, and Qin Jin. 2022. Rucaim3-tencent at trecvid 2022: Video to text description. In *Proceedings of TRECVID*.

Liang Zhang, Yuqing Song, and Qin Jin. 2021a. Ruc_aim3 at trecvid 2021: Video to text description. In *Proceedings of TRECVID*.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021b. Vinvl: Revisiting visual representations in vision-language models. In *Proc. of CVPR*, pages 5579–5588.