T2V2T: Text-to-Video-to-Text Fusion for Text-to-Video Retrieval

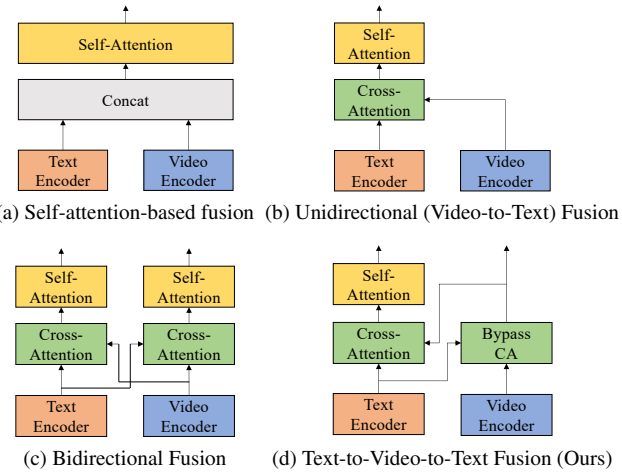
Jonghee Kim[1] Youngwan Lee[1,2] Jinyoung Moon[1,3]
[1]Electronics and Telecommunications Research Institute (ETRI), South Korea
[2]Korea Advanced Institute of Science and Technology (KAIST), South Korea
[3]University of Science and Technology (UST), South Korea
{jhkim27, yw.lee, jymooon}@etri.re.kr

Abstract

Video-language transformers for text-to-video retrieval typically consist of a video encoder, a text encoder, and a joint encoder. The joint encoder can be categorized into 1) self-attention-based fusion and 2) unidirectional fusion based on cross-attention. The former approach performs self-attention on the concatenation of video and text embeddings. Although it allows complete interaction between text and video, the length of the input sequences makes it computationally intensive. Instead, unidirectional fusion employs rather efficient cross-attention to fuse video embeddings into text embeddings while ignoring text-to-video interaction. The text-to-video fusion is not well explored because of the information imbalance between text and video, which makes it difficult to determine which video patches can be used as queries in cross-attention. i.e., a text embedding corresponds to one or more patch embeddings, while a video patch embedding may not correspond to any text embeddings. In order to address this challenge, we devise a Bypass cross-attention (Bypass CA) which prevents matching between irrelevant video and text embedding pairs in the cross-attention. Using Bypass CA, we propose a novel bidirectional interaction approach, Text-to-Video-to-Text (T2V2T) fusion. The proposed T2V2T uses two unidirectional fusions with opposite directions, i.e.*, text-to-video fusion followed by video-to-text fusion. As a result, the proposed T2V2T fusion yields state-of-the-art results on MSR-VTT, DiDeMo, and ActivityNet Captions.*

1. Introduction

Recently, text-to-video retrieval methods have leveraged video-language pre-training (VLP) [2, 5, 7, 9, 12, 17, 24–26, 33] since VLP has shown superior performance in downstream tasks such as text-to-video retrieval [1, 11, 28], video question answering [14, 27, 30, 32], and video captioning [11, 28, 34]. Video-language models (VLMs) for VLP commonly employ a text encoder, a video encoder, and a joint encoder. Whether or not the joint encoder is employed, we could categorize VLMs into dual and joint encoder-based models. Dual encoder-based models [2, 9, 24] employ text and video encoders to learn shared representation using contrastive learning [15]. Joint encoder-based models [5, 7, 12, 17, 26, 33] embed each modality into video and text embeddings separately and then merge the embeddings with a joint encoder for downstream tasks. Although dual encoder-based models are efficient, they performed worse than joint encoder-based models in a variety of tasks due to the lack of video-language *interaction*.



(a) Self-attention-based fusion (b) Unidirectional (Video-to-Text) Fusion
(c) Bidirectional Fusion (d) Text-to-Video-to-Text Fusion (Ours)

Figure 1. Examples of joint encoder configuration.

Existing joint encoder-based models for fusing video and text could be categorized into self-attention-based fusion (Fig. 1a) and unidirectional fusion based on cross-attention (Fig. 1b). Since self-attention-based approaches [7, 17, 25, 33] take concatenated video and text embeddings as input, they require a huge computational cost since the computational cost is proportional to the square of the input sequence length. In order to reduce the computational cost, unidirectional fusion methods [2, 5, 12, 26] fuse video embeddings into text embeddings by using cross-attention followed by self-attention on the fused text em-

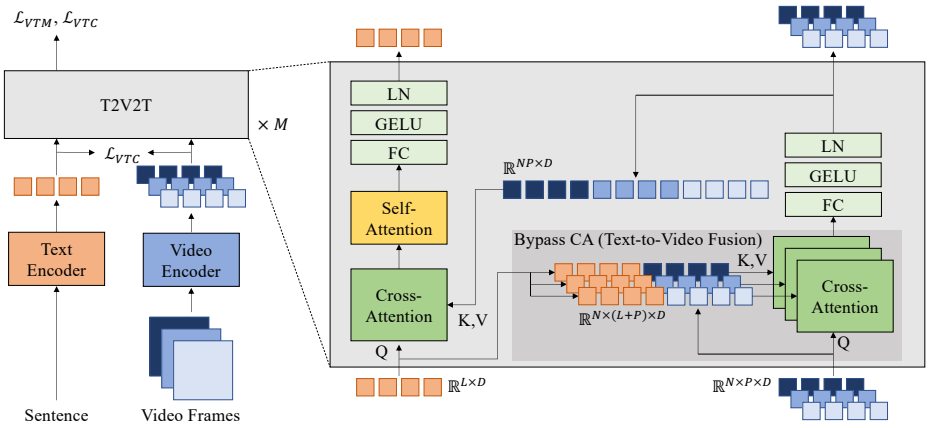


Figure 2. Overall architecture. Text embeddings and video embeddings are given to T2V2T fusion. In T2V2T fusion, video embeddings are first modulated using both video and text embeddings in Bypass CA. And then, the modulated video embeddings are fused into text embeddings using cross-attention.

beddings as shown in Fig. 1b. Although applying self-attention with only text embeddings is efficient, it results in limited unidirectional interaction, *i.e.*, video-to-text interaction without text-to-video interaction. As a straightforward extension of unidirectional fusion, Cheng *et al.* [5] explore a bidirectional fusion method (Fig. 1c), where two cross-attentions transfer each video and text embedding to each other. However, the bidirectional method shows similar performance even with extra text-to-video fusion. We assume that it is due to the property of cross-attention in text-to-video fusion, which associates all patch embeddings as queries with the text embeddings, regardless of the correlation between each patch and the text embedding. For example, the number of text tokens is less than 100 in our experiments, while the number of patch tokens is up to 6,272 for 32 frames. Due to the imbalance, a text token can correspond to one or more patch tokens, while a patch token may not correspond to any text tokens. It could cause improper modification of patch tokens by using irrelevant text tokens.

To resolve this limitation, we devise Bypass cross-attention (Bypass CA), which prevents video patch embeddings from being modified by the use of irrelevant text. To this end, we copy query video embeddings into the key and value in addition to text features in text-to-video cross-attention to make video embeddings associate with themselves in the key instead of text features if the given sentence and each frame are not relevant to each other. With Bypass CA, we propose an effective bidirectional fusion method (Fig. 1d), called Text-to-Video-to-Text (T2V2T) fusion, by enabling text-to-video interaction. Concretely, T2V2T fusion first transfers text embeddings to video embeddings using Bypass CA and then transfers the modulated video embeddings back to text embeddings using cross-attention. For video-to-text fusion, we follow the previous unidirectional fusion methods [5, 12]. The video-fused text features are then re-integrated by self-attention. The T2V2T fusion is performed recursively for iterative interaction.

Our main contributions are summarized as follows. 1) We investigate the text-to-video interaction, which suffers from the imbalance between the number of video and text embeddings. 2) We propose a novel fusion method called T2V2T fusion by introducing Bypass CA. 3) T2V2T achieves SOTA text-to-video retrieval results on MSR-VTT [28], DiDeMo [1], and ActivityNet Captions [11].

2. Proposed Method

The proposed text-to-video retrieval follows previous VLP-based approaches, *i.e.*, pre-training followed by fine-tuning on text-to-video retrieval. In this section, we describe our VLM with T2V2T fusion based on Bypass CA, training objectives and implementation details.

Overall Architecture is shown in Fig. 2. It consists of a text encoder, a video encoder, and a joint encoder (T2V2T fusion). The text encoder takes a sentence of length L as input and yields text features $F_T \in \mathbb{R}^{L\times D}$, where D is an embedding dimension. The video encoder takes N frames of a video as input and returns video features $F_V \in \mathbb{R}^{N\times P\times D}$, where P is the number of patches in a frame. The joint encoder (T2V2T fusion) takes F_T and F_V as input and outputs video-text features. The video-text features are employed for text-to-video-retrieval through a fully connected layer.

Text-to-Video-to-Text (T2V2T) Fusion is a bidirectional fusion encoder that extends the unidirectional encoder employed in [5, 12]. In [5, 12], a cross-attention is employed to fuse video features into text features without regard to text-to-video interaction. In order to enable text-to-video interaction, we introduce Bypass cross-attention (Bypass CA) in addition to the unidirectional encoder.

Since cross-attention in the text-to-video fusion of bidirectional fusion [5] (Fig. 1c) associates all frames with the given sentence, video features in each frame are modulated

Method	#PT	MSRVTT				DiDeMo				ActivityNet Captions			
		R1	R5	R10	Avg.	R1	R5	R10	Avg.	R1	R5	R10	Avg.
ClipBERT [13]	5.6M	22.0	46.8	59.9	42.9	20.4	48.0	60.8	43.1	21.3	49.0	63.5	44.6
Frozen [2]	5.5M	31.0	59.5	70.5	53.7	31.0	59.8	72.4	54.4	-	-	-	-
ALPRO [15]	5.5M	33.9	60.7	73.2	55.9	35.9	67.5	78.8	60.7	-	-	-	-
BridgeFormer [9]	5.5M	37.6	64.8	75.1	59.2	37.0	62.2	73.9	57.7	-	-	-	-
Singularity [12]	5.5M	39.9	67.3	76.0	61.1	49.2	77.5	85.4	70.7	45.9	73.3	83.8	67.7
VindLU [5]	5.5M	43.8	70.3	**79.5**	64.5	54.6	81.3	89.0	75.0	51.1	79.2	**88.4**	72.9
T2V2T (Ours)	5.5M	**44.4**	**70.7**	**79.5**	**64.9**	**56.0**	**81.9**	**89.7**	**75.9**	**52.1**	**79.4**	88.2	**73.2**
MMT [8]	136M	25.8	57.2	69.3	50.8	-	-	-	-	28.7	61.4	94.5	61.5
TACo [29]	120M	28.4	57.8	71.2	52.5	-	-	-	-	30.4	61.2	93.4	61.7
SupportSet [22]	120M	30.1	58.5	69.3	52.6	-	-	-	-	29.2	61.6	94.7	61.8
Singularity [12]	17M	42.7	69.5	78.1	63.4	53.1	79.9	88.1	73.7	48.9	77.0	86.3	70.7
VindLU [5]	17M	45.3	69.9	79.6	64.9	59.2	84.1	89.5	77.6	54.4	80.7	89.0	74.7
CLIP4Clip [21]	400M	44.5	71.4	81.6	65.8	43.4	70.2	80.6	64.7	40.5	72.4	98.2	70.4
VindLU [5]	25M	46.5	71.5	80.4	66.1	61.2	85.8	91.0	79.3	55.0	81.4	89.7	75.4
OmniVL [26]	17M	47.8	74.2	83.8	68.6	52.4	79.5	85.4	72.4	-	-	-	-

Table 1. Comparison to state-of-the-art VLP-based methods on text-to-video retrieval. #PT is the number of images and videos used for pre-training. R1, R5, and R10 are recall at 1, 5, and 10, respectively. We also report results obtained by pre-training with $\geq$ 17M data as a reference, but they are grayed out for a fair comparison.

by text features without regard to the correlation between the given sentence and each frame. We assume that it is inappropriate to associate all frames with the given sentence since only a subset of the frames is relevant to the given sentence. For example, self-attention-based fusion Fig. 1a can discourage the association between irrelevant words and frames since self-attention is performed using both embeddings. As an efficient alternative, we introduce a bypass mechanism in the cross-attention by including frame features in the key and value in addition to text features, so that frame features (*i.e.*, query) can be associated with themselves in the key instead of text features if the given sentence and each frame are not relevant to each other as follows:

$$F_{T2V}[i] = \texttt{X-Attn}(F_V[i], F_T||F_V[i]), \quad (1)$$

where $F_V[i]$ is video features of i-th frame. $\texttt{X-Attn}(Q, KV)$ is a cross-attention where Q is a query, KV is employed as key and value. Note that Bypass CA is performed in a frame-wise manner to avoid a significant increase in computational cost.

Then, F_{T2V} is fused into text features as follows:

$$F_{T2V2T} = \texttt{S-Attn}(\texttt{X-Attn}(F_T, F_{T2V})), \quad (2)$$

where $\texttt{S-Attn}(\cdot)$ is a self-attention to further refine the fused features. Note that we omit a fully connected layer, GELU [10] activation function, and a layer normalization layer after Eq. (1) and Eq. (2) for simplicity. T2V2T fusion is performed M times recursively for iterative interaction.

Objectives In the pre-training phase, we employ three common objectives; Video-Text Contrastive loss ($\mathcal{L}_{VTC}$) [15] on text and video embeddings, Video-Text Matching loss ($\mathcal{L}_{VTM}$) [15, 16, 20], and Masked Language Modeling loss ($\mathcal{L}_{MLM}$) [6]. For finetuning on text-to-video retrieval, we employ $\mathcal{L}_{VTC}$ and $\mathcal{L}_{VTM}$ except $\mathcal{L}_{MLM}$.

Implementation details. We follow the setup of [5] to fairly compare proposed T2V2T fusion encoder with the state-of-the-art unidirectional encoder [5]. The video encoder is based on BEiT [3], and temporal attention inspired by TimeSformer [4] is inserted before each spatial attention. The first nine layers of BERT-Base [6] are employed for the text encoder, and the last three layers are employed to initialize self-attention in the T2V2T fusion encoder. We randomly initialize the temporal attention layers in the video encoder and the cross-attention layers in the T2V2T fusion.

3. Experiments

In this section, we first describe experiments setup of pre-training (Sec. 3.1) and text-to-video retrieval (Sec. 3.2), then demonstrate experimental results compared to state-of-the-art VLP-based text-to-retrieval methods [2, 5, 9, 12, 13, 15] in Sec. 3.3. In addition, we show an ablation study to verify the effectiveness of Bypass CA in Sec. 3.4.

3.1. Pre-training

We pre-trained our model on CC3M [23] and WebVid-2M [2], which are common for VLP. For a fair comparison, we followed the training setting in [5] as shown in Tab. 3.

Method	MSR-VTT				DiDeMo				ActivityNet Captions				Total
	R1	R5	R10	Avg.	R1	R5	R10	Avg.	R1	R5	R10	Avg.	Avg.
VindLU [5]	43.8	70.3	**79.5**	64.5	54.6	81.3	89.0	75.0	51.1	79.2	**88.4**	72.9	70.8
naïve T2V2T	44.3	70.1	79.3	64.6	55.1	80.7	88.0	74.6	51.7	78.8	87.9	72.8	70.7
T2V2T	**44.4**	**70.7**	**79.5**	**64.9**	**56.0**	**81.9**	**89.7**	**75.9**	**52.1**	**79.4**	88.2	**73.2**	**71.3**

Table 2. An ablation study to verify the efficacy of the proposed Bypass CA.

config	parameters
optimizer	AdamW [19] ($\beta_1 = 0.9$, $\beta_2 = 0.999$, wd=0.02)
learning rate	1e-4→1e-6 (cosine decay [18])
#epochs	10 (warmup = 1)
batch size×#GPUs	64×8
spatial resolution	224 × 224
Augmentation	random resize, crop horizontal flip
#training frames	4

Table 3. **Hyper-parameters for pre-training.**

config	parameters		
	MSR-VTT	DiDeMo	Anet Cap.
learning rate	1e-5→1e-6 (cosine decay [18])		
#epochs	5	10	10
	(warmup = 0.5)		
batch size×#GPUs	32×4	32×1	32×1
#training frames	12	12	12
#inference frames	12	12	32

Table 4. **Hyper-parameters for text-to-video retrieval.** We omit parameters which are common in the pre-training (Tab. 3).

3.2. Text-to-video retrieval

We evaluated VLP-based text-to-video retrieval methods on the following three datasets:

- **MSR-VTT** [28] contains 10K videos with 200K captions. We trained on 9K videos and evaluated on 1K-A test set following [2, 5, 31].
- **DiDeMo** [1] contains 10K videos with 41K captions. We trained on training set with 8,395 videos and evaluated on test set with 1,004 videos following [5, 13, 21].
- **ActivityNet Captions** [11] contains 20K videos with 100K captions. We trained on training set with 10K videos and evaluated on validation set with 4.9K videos following [5, 12, 21]

We followed the training setting of [5] as shown in Tab. 4 for a fair comparison.

3.3. Experimental Results

We compared the proposed method with state-of-the-art VLP-based methods in terms of text-to-video retrieval recall as shown in Tab. 1. Since the number of images and videos used in the pre-training phase affects the performance, we also report the number of images and videos. Our proposed method achieved the best recall under 5.5M pre-training data. Notably, the proposed method obtained gains in recall at 1, +0.6 on MSR-VTT, +1.4 on DiDeMo, and +1.0 on ActivityNet Captions compared to the second best method, VindLU (5.5M) [5]. In addition, the proposed method achieved better results than the methods pre-trained on 17-400M data except VindLU (17M/25M) [5] for DiDeMo and ActivityNet Captions text-to-video retrieval.

3.4. Ablation Study

In order to verify the effectiveness of the proposed T2V2T and Bypass CA, we compared three methods in terms of text-to-video retrieval recall on three text-to-video retrieval datasets; baseline unidirectional fusion [5], naïve T2V2T ([5] + cross-attention), and our T2V2T ([5] + Bypass CA). As shown in Tab. 2, naïve T2V2T yielded slightly worse results than its baseline unidirectional fusion. It is the same phenomenon as [5] that bidirectional fusion does not improve over unidirectional fusion. We speculate that it is due to the property of cross-attention in text-to-video fusion, which associates irrelevant frames and text. In contrast, the proposed Bypass CA improved over unidirectional fusion by taking advantage of the bypass mechanism.

4. Conclusion

In this paper, we have proposed Bypass cross-attention to deal with the information imbalance between video and text. Using Bypass cross-attention, the proposed T2V2T fusion achieves state-of-the-art results in text-to-video retrieval. We will extend this work to a variety of downstream tasks to verify the generality of Bypass CA and scale up the pre-training data for further performance gains.

5. Acknowledgement

This work was supported by IITP grant funded by the Korea government (MSIT) (No. 2020-0-00004, Development of Previsional Intelligence based on Long-term Visual Memory Network (80%), and No. RS-2022-00187238, Development of Large Korean Language Model Technology for Efficient Pre-training (20%)).

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, 4

[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 (ICCV)*, pages 1708–1718, 2021. 1, 3, 4

[3] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. In *International Conference on Learning Representations*, 2022. 3

[4] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021. 3

[5] Feng Cheng, Xizi Wang, Jie Lei, David Crandall, Mohit Bansal, and Gedas Bertasius. Vindlu: A recipe for effective video-and-language pretraining. *arXiv preprint arXiv:2212.05051*, 2022. 1, 2, 3, 4

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. 3

[7] Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. Violet: End-to-end video-language transformers with masked visual-token modeling. *arXiv preprint arXiv:2111.12681*, 2021. 1

[8] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 214–229, 2020. 3

[9] Yuying Ge, Yixiao Ge, Xihui Liu, Dian Li, Ying Shan, Xiaohu Qie, and Ping Luo. Bridging video-text retrieval with multiple choice questions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 3

[10] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 3

[11] 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, 2, 4

[12] Jie Lei, Tamara L Berg, and Mohit Bansal. Revealing single frame bias for video-and-language learning. *arXiv*, 2022. 1, 2, 3, 4

[13] 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 (CVPR)*, 2021. 3, 4

[14] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. In *Empirical Methods in Natural Language Processing*, 2018. 1

[15] Dongxu Li, Junnan Li, Hongdong Li, Juan Carlos Niebles, and Steven C. H. Hoi. Align and prompt: Video-and-language pre-training with entity prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4953–4963, 2022. 1, 3

[16] Linjie Li, Yen Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+language omni-representation pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2046–2065, 2020. 3

[17] Linjie Li, Zhe Gan, Kevin Lin, Chung-Ching Lin, Zicheng Liu, Ce Liu, and Lijuan Wang. Lavender: Unifying video-language understanding as masked language modeling. *arXiv preprint arXiv:2206.07160*, 2022. 1

[18] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017. 4

[19] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 4

[20] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*, 2020. 3

[21] 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. *arXiv*, 2021. 3, 4

[22] Mandela Patrick, Po-Yao Huang, Yuki Asano, Florian Metze, Alexander G Hauptmann, Joao F Henriques, and Andrea Vedaldi. Support-set bottlenecks for video-text representation learning. In *International Conference on Learning Representations*, 2021. 3

[23] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 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, 2018. 3

[24] Fangxun Shu, Biaolong Chen, Yue Liao, Shuwen Xiao, Wenyu Sun, Xiaobo Li, Yousong Zhu, Jinqiao Wang, and Si Liu. Masked contrastive pre-training for efficient video-text retrieval. *arXiv preprint arXiv:2212.00986*, 2022. 1

[25] Alex Jinpeng Wang, Yixiao Ge, Rui Yan, Yuying Ge, Xudong Lin, Guanyu Cai, Jianping Wu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. All in one: Exploring unified video-language pre-training. *arXiv preprint arXiv:2203.07303*, 2022. 1

[26] Junke Wang, Dongdong Chen, Zuxuan Wu, Chong Luo, Luowei Zhou, Yucheng Zhao, Yujia Xie, Ce Liu, Yu-Gang Jiang, and Lu Yuan. Omnivl: One foundation model for image-language and video-language tasks. In *Advances in Neural Information Processing Systems*, 2022. 1, 3

[27] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653, 2017. 1

[28] 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, 2, 4

[29] Jianwei Yang, Yonatan Bisk, and Jianfeng Gao. Taco: Token-aware cascade contrastive learning for video-text alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11562–11572, 2021. 3

[30] Youngjae Yu, Jongseok Kim, and Gunhee Kim. A joint sequence fusion model for video question answering and retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 471–487, 2018. 1

[31] Youngjae Yu, Jongseok Kim, and Gunhee Kim. A joint sequence fusion model for video question answering and retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 471–487, 2018. 4

[32] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9127–9134, 2019. 1

[33] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. *Advances in Neural Information Processing Systems*, 34:23634–23651, 2021. 1

[34] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 1