Cap4Video: What Can Auxiliary Captions Do for Text-Video Retrieval?

Wenhao Wu[1,2] *Haipeng Luo*[3] *Bo Fang*[3] *Jingdong Wang*[2] *Wanli Ouyang*[4,1]

[1]The University of Sydney [2]Baidu Inc. [3]UCAS [4]Shanghai AI Laboratory

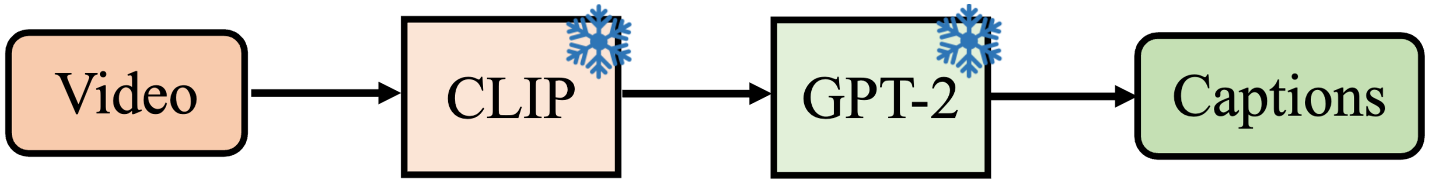
JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Code & Models

WED-AM-236

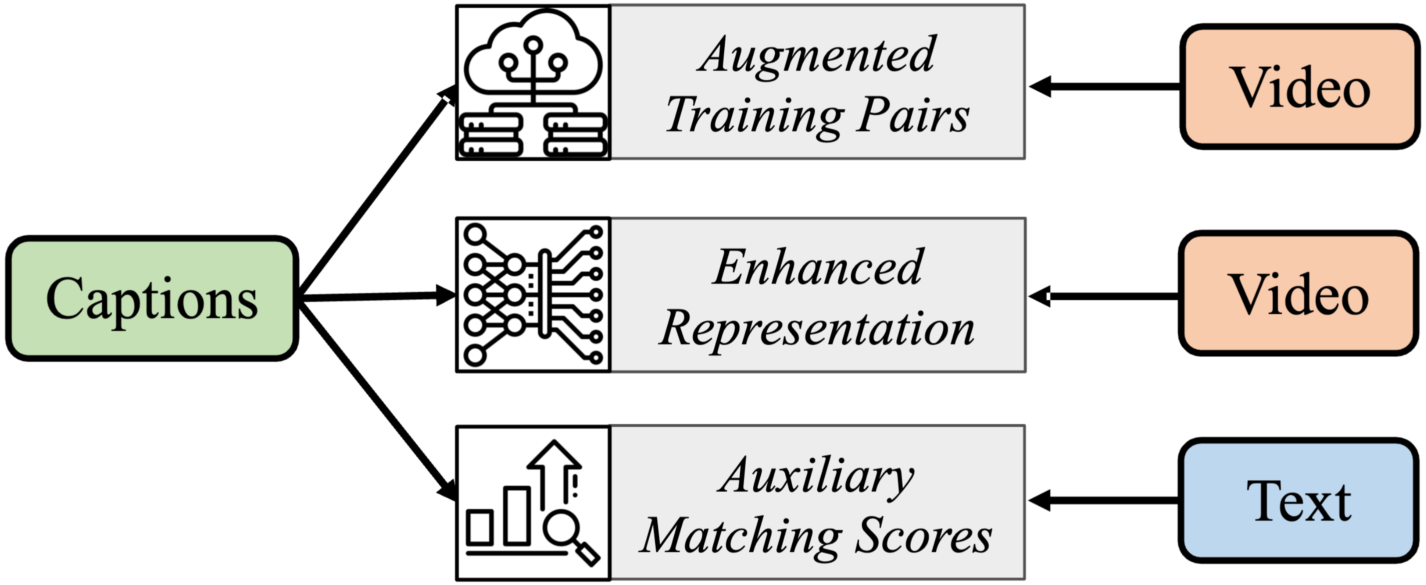
Key Innovation

(a) Zero-Shot Caption Generation



We leverage the knowledge of large-scale vision-language models (VLMs), such as CLIP, and large language models (LLMs), such as GPT-2, to generate diverse captions for arbitrary videos.

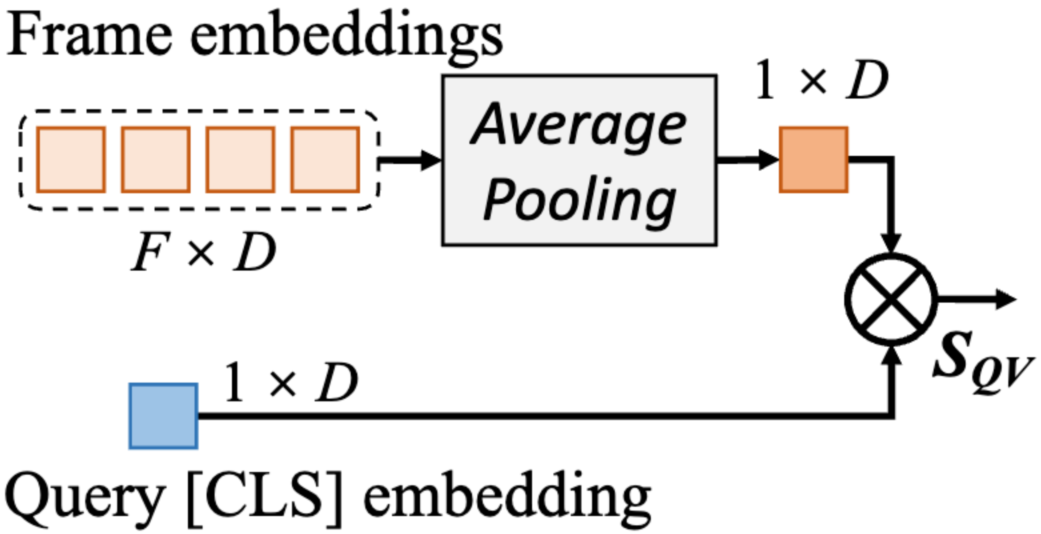
(b) Captions for Text-Video Matching



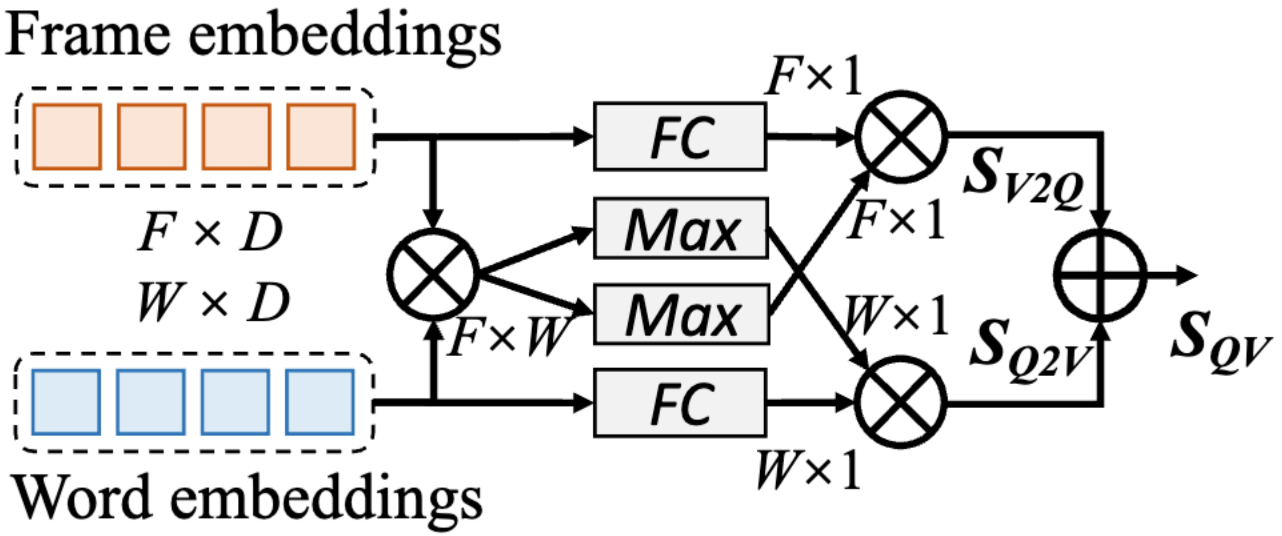
Our **Cap4Video** improves upon existing text-video retrieval methods through three key aspects.

Background: Text-Video Retrieval

① Global embedding matching

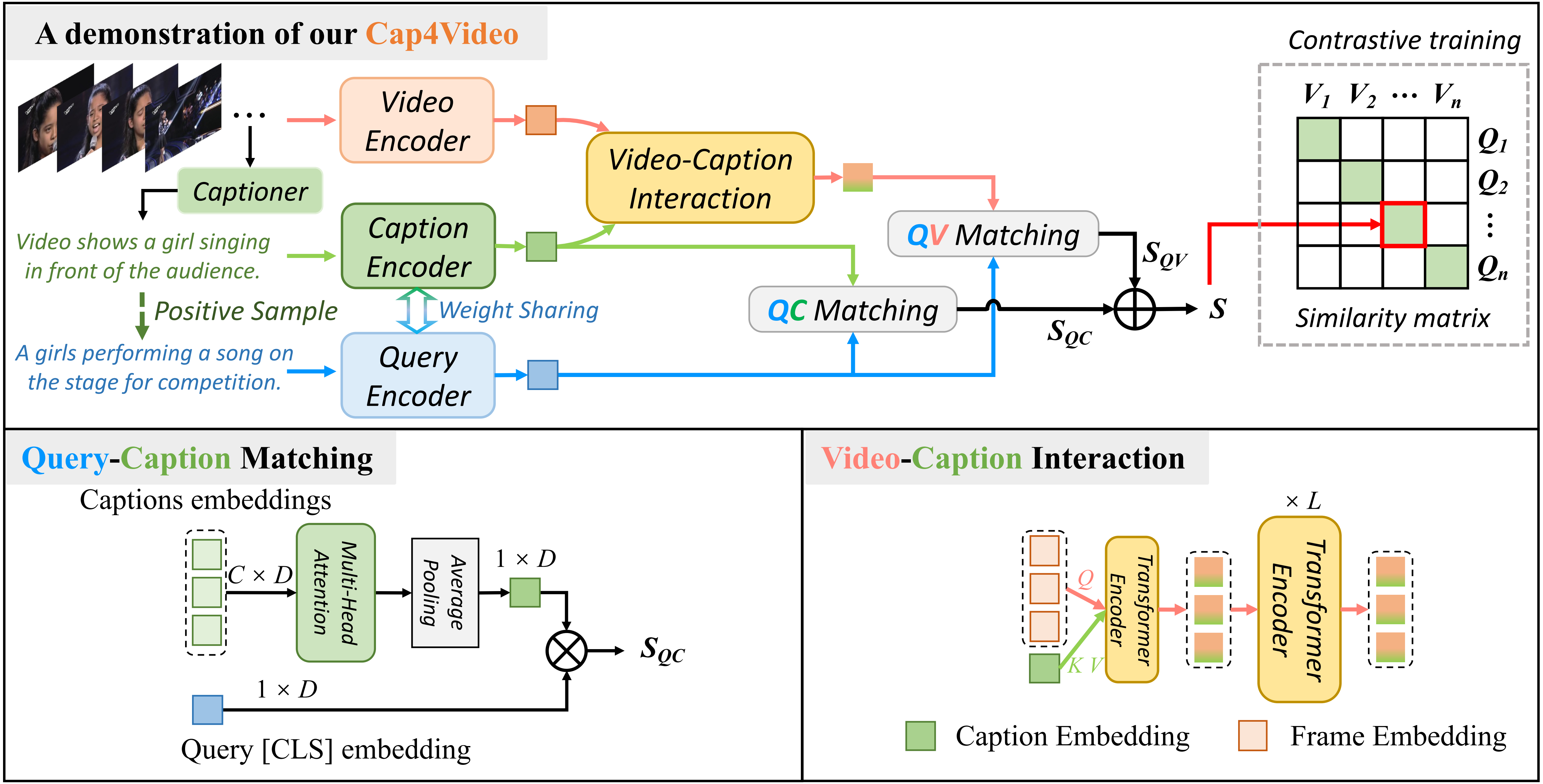


② Fine-grained embedding matching



Query-Video Matching: Two typical mechanisms

Our Method


A demonstration of our Cap4Video
Contrastive training
Video Encoder
Captioner
Video shows a girl singing in front of the audience.
Caption Encoder
Positive Sample
Weight Sharing
A girls performing a song on the stage for competition.
Query Encoder
Video-Caption Interaction
QV Matching
QC Matching
S_{QV}
S_{QC}
S
V_1 V_2 … V_n
Q_1
Q_2
Q_n
Similarity matrix
Query-Caption Matching
Captions embeddings
$C \times D$
Multi-Head Attention
Average Pooling
$1 \times D$
S_{QC}
$1 \times D$
Query [CLS] embedding
Video-Caption Interaction
$\times L$
Q
K V
Transformer Encoder
Transformer Encoder
Caption Embedding
Frame Embedding

Learning Objectives

Query-Caption Matching Loss

$$\mathcal{L}_{Q2C} = -\frac{1}{B}\sum_{i}^{B} \log \frac{\exp(s_{qc}(\mathbf{e_t}_i, \mathbf{e_c}_i)/\tau)}{\sum_j^B \exp(s_{qc}(\mathbf{e_t}_i, \mathbf{e_c}_j)/\tau)},$$

$$\mathcal{L}_{C2Q} = -\frac{1}{B}\sum_{i}^{B} \log \frac{\exp(s_{qc}(\mathbf{e_t}_i, \mathbf{e_c}_i)/\tau)}{\sum_j^B \exp(s_{qc}(\mathbf{e_t}_j, \mathbf{e_c}_i)/\tau)},$$

$$\mathcal{L}_{QC} = \frac{1}{2}(\mathcal{L}_{Q2C} + \mathcal{L}_{C2Q}),$$

Query-Video Matching Loss

$$\mathcal{L}_{Q2V} = -\frac{1}{B}\sum_{i}^{B} \log \frac{\exp(s_{qv}(\mathbf{e_t}_i, \mathbf{e_v}_i)/\tau)}{\sum_j^B \exp(s_{qv}(\mathbf{e_t}_i, \mathbf{e_v}_j)/\tau)},$$

$$\mathcal{L}_{V2Q} = -\frac{1}{B}\sum_{i}^{B} \log \frac{\exp(s_{qv}(\mathbf{e_t}_i, \mathbf{e_v}_i)/\tau)}{\sum_j^B \exp(s_{qv}(\mathbf{e_t}_j, \mathbf{e_v}_i)/\tau)},$$

$$\mathcal{L}_{QV} = \frac{1}{2}(\mathcal{L}_{Q2V} + \mathcal{L}_{V2Q}),$$

Total Loss

$$\mathcal{L} = \mathcal{L}_{QV} + \mathcal{L}_{QC}.$$

Experiments

- Experimental results:
  - Comparison to the state-of-the-art methods on text-video retrieval.

- Datasets:
  - **MSR-VTT**: ~10K video videos, each having 20 captions;
  - **DiDeMo**: ~10K videos paired with 40K description;
  - **VATEX**: ~35K videos, each with multiple annotations;
  - **MSVD**: 1970 videos with 80K captions, with ~40 captions on average per video.

Comparisons with SOTAs

Method	Venue	Text → Video					Video → Text				
		R@1	R@5	R@10	MdR↓	MnR↓	R@1	R@5	R@10	MdR↓	MnR↓
ClipBERT [19]	CVPR'20	22.0	46.8	59.9	6.0	-	-	-	-	-	
MMT [10]	ECCV'20	26.6	57.1	69.6	4.0	-	27.0	57.5	69.7	3.7	21.3
T2VLAD [39]	CVPR'21	29.5	59.0	70.1	4.0	-	31.8	60.0	71.1	3.0	
SupportSet [29]	ICLR'21	30.1	58.5	69.3	3.0	-	28.5	58.6	71.6	3.0	-
Frozen [2]	ICCV'21	32.5	61.5	71.2	3.0	-	-	-	-	-	-
BridgeFormer [12]	CVPR'22	37.6	64.8	75.1	-	-	-	-	-	-	-
TMVM [20]	NeurIPS'22	36.2	64.2	75.7	3.0	-	34.8	63.8	73.7	3.0	-
CLIP-ViT-B/32											
CLIP4Clip [24]	arXiv'21	44.5	71.4	81.6	2.0	15.3	42.7	70.9	80.6	2.0	11.6
CenterCLIP [48]	SIGIR'22	44.2	71.6	82.1	2.0	15.1	42.8	71.7	82.2	2.0	10.9
CAMoE [7]	arXiv'21	44.6	72.6	81.8	2.0	13.3	45.1	72.4	83.1	2.0	10.0
CLIP2Video [9]	arXiv'21	45.6	72.6	81.7	2.0	14.6	43.5	72.3	82.1	2.0	10.2
X-Pool [13]	CVPR'22	46.9	72.8	82.2	2.0	14.3	-	-	-	-	-
QB-Norm [4]	CVPR'22	47.2	73.0	83.0	2.0	-	-	-	-	-	-
TS2-Net [22]	ECCV'22	47.0	74.5	83.8	2.0	13.0	45.3	74.1	83.7	2.0	9.2
DRL [36]	arXiv'22	47.4	74.6	83.8	2.0	-	45.3	73.9	83.3	2.0	-
Cap4Video		**49.3**	74.3	83.8	2.0	12.0	**47.1**	73.7	84.3	2.0	8.7
CLIP-ViT-B/16											
CLIP2TV [11]	arXiv'21	48.3	74.6	82.8	2.0	14.9	46.5	75.4	84.9	2.0	10.2
CenterCLIP [48]	SIGIR'22	48.4	73.8	82.0	2.0	13.8	47.7	75.0	83.3	2.0	10.2
TS2-Net [22]	ECCV'22	49.4	75.6	85.3	2.0	13.5	46.6	75.9	84.9	2.0	8.9
DRL [36]	arXiv'22	50.2	76.5	84.7	1.0	-	48.9	76.3	85.4	2.0	-
Cap4Video		**51.4**	75.7	83.9	1.0	12.4	**49.0**	75.2	85.0	2.0	8

Results on MSR-VTT 1K dataset

Method	R@1	R@5	R@10	MdR	MnR
CE [21]	15.6	40.9	-	8.2	-
ClipBERT [19]	21.1	47.3	61.1	6.3	-
Frozen [2]	31.0	59.8	72.4	3.0	-
TMVM [20]	36.5	64.9	75.4	3.0	-
CLIP4Clip [24]	42.8	68.5	79.2	2.0	18.9
TS2-Net [22]	41.8	71.6	82.0	2.0	14.8
HunYuan [28]	45.0	75.6	83.4	2.0	12.0
DRL [36]	49.0	76.5	84.5	2.0	-
Cap4Video	**52.0**	79.4	87.5	1	10.5

Results on DiDeMo dataset

Method	R@1	R@5	R@10	MdR	MnR
CE [21]	19.8	49.0	63.8	6.0	-
SUPPORT [29]	28.4	60.0	72.9	4.0	-
CLIP [30]	37.0	64.1	73.8	3.0	-
Frozen [2]	33.7	64.7	76.3	3.0	-
TMVM [20]	36.7	67.4	81.3	2.5	-
CLIP4Clip [24]	45.2	75.5	84.3	2.0	10.3
X-Pool [13]	47.2	77.4	86.0	2.0	9.3
Cap4Video	**51.8**	80.8	88.3	1	8.3

Results on MSVD dataset

Method	R@1	R@5	R@10	MdR	MnR
HGR [6]	35.1	73.5	83.5	2.0	-
CLIP [30]	39.7	72.3	82.2	2.0	12.8
SUPPORT [29]	44.9	82.1	89.7	1.0	-
CLIP4Clip [24]	55.9	89.2	95.0	1.0	3.9
Clip2Video [9]	57.3	90.0	95.5	1.0	3.6
QB-Norm [4]	58.8	88.3	93.8	1.0	-
TS2-Net [22]	59.1	90.0	95.2	1.0	3.5
Cap4Video	**66.6**	93.1	97.0	1	2.7

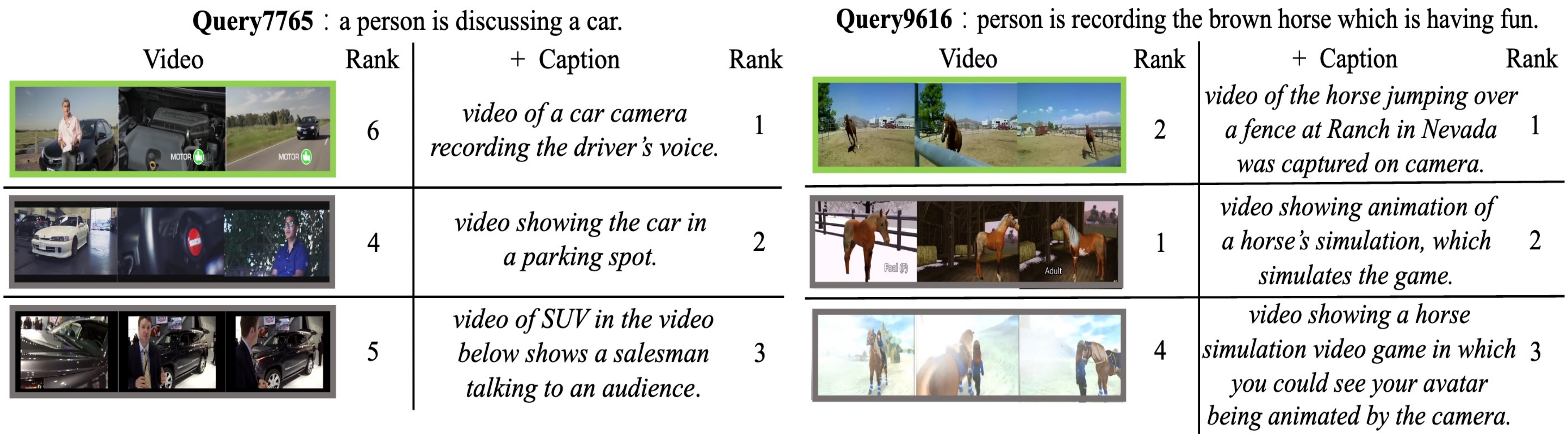
Results on VATEX dataset

Ablation Studies

Method	Global Matching					Fine-grained Matching				
	R@1	R@5	R@10	MdR↓	MnR↓	R@1	R@5	R@10	MdR↓	MnR↓
Baseline	42.8	70.4	79.0	2	16.6	45.7	73.7	82.6	2	13.1
+Different Sources of Caption as Data Augmentation										
Video Title from Source URL	43.8	71.1	80.9	2	15.1	44.3	72.7	83.5	2	13.1
Zero-shot Video Captioning	**44.2**	70.7	81.5	2	16.2	**46.3**	72.5	81.7	2	12.9
+Different Number of Captions for Data Augmentation										
Top-1	**44.2**	70.7	81.5	2	16.2	**46.3**	72.5	81.7	2	12.9
Top-3	43.3	71.7	81.6	2	15.0	45.5	73.8	82.4	2	12.7
Top-5	43.4	70.6	80.4	2	16.2	45.6	72.7	82.7	2	12.9
+Video-Caption Feature Interaction										
Video Only	44.2	70.7	81.5	2	16.2	46.3	72.5	81.7	2	12.9
Sum	43.8	71.5	80.3	2	16.1	47.2	73.3	82.8	2	13.1
Concat-MLP	37.5	66.1	78.4	3	15.7	40.0	68.7	79.9	2	12.7
Cross Transformer	44.6	71.6	80.3	2	14.6	47.9	75.4	83.0	2	11.5
Co-attention Transformer	**45.3**	71.2	80.9	2	15.0	**48.5**	74.0	82.5	2	12.7
+Query-Caption Matching Score										
Query-Video Only	45.3	71.2	80.9	2	15.0	48.5	74.0	82.5	2	12.7
Query-Caption Only	30.3	55.2	67.5	4	26.4	30.3	55.2	67.5	4	26.4
Query-Video + Query-Caption	**45.6**	71.7	81.2	2	14.8	**49.3**	74.2	83.4	2	12.1

Component-wise evaluation of our framework on the MSR-VTT 1K validation set.

Visualization



The text-video results on the MSR-VTT 1K-A test set. **Left**: The ranking results of the query-video matching model. **Right**: The ranking results of Cap4Video, which incorporates generated captions to enhance retrieval.

Conclusion

- We explore a novel problem: leveraging auxiliary captions to further enhance existing text-video retrieval.

- We propose the **Cap4Video**, which maximizes the utility of the auxiliary captions through **three** aspects: 1) Input data augmentation for training, 2) Intermediate video-caption feature interaction for compact video representations, and 3) Output score fusion for improved text-video retrieval.

- Our Cap4Video improves the performance of existing query-video matching mechanisms, including global matching and fine-grained matching. It has achieved state-of-the-art performance across four standard text-video retrieval benchmarks.

THANKS

🔥 Codes & Models

https://github.com/whwu95/Cap4Video

👫 Contact

Wenhao Wu

Email: whwu.ucas@gmail.com

Homepage:

https://whwu95.github.io