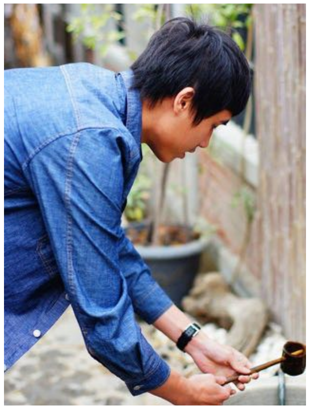
VSLab

Microsoft Research 微软亚洲研究院

科 技 部 Ministry of Science and Technology

NOVATEK

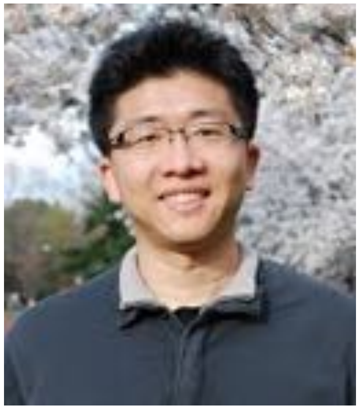
Video Title Generation



Kuo-Hao Zeng
NTHU EE

Tseng-Hung Chen
NTHU EE

Juan Carlos Niebles
Stanford CS

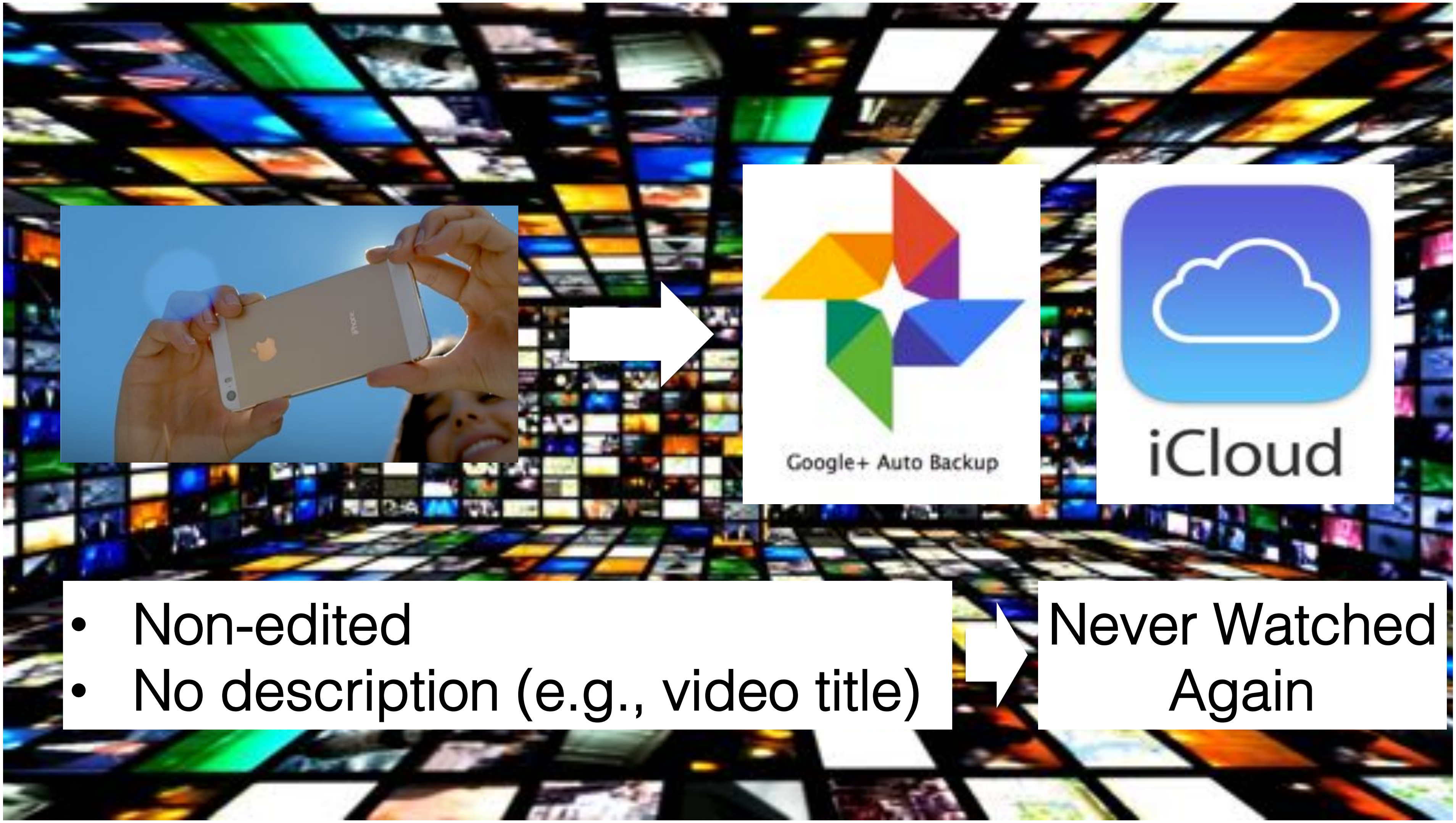


Min Sun
NTHU EE

Present at ECCV'16 EUROPEAN CONFERENCE ON COMPUTER VISION

October 8 – 16, 2016 | Amsterdam | the Netherlands

Motivation



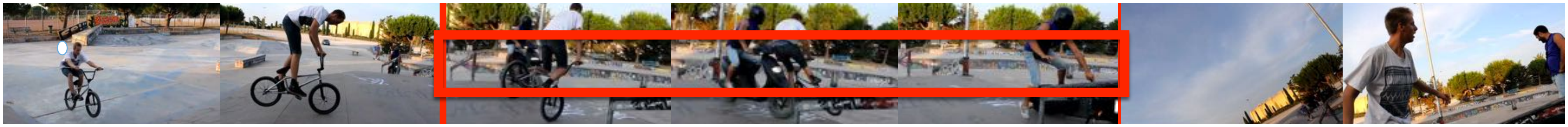
- Non-edited
- No description (e.g., video title)

Never Watched Again

VSLab

What If?

- Detect the highlight moment



- Generate a description of the highlight

Bmx rider gets *hit by scooter* at park

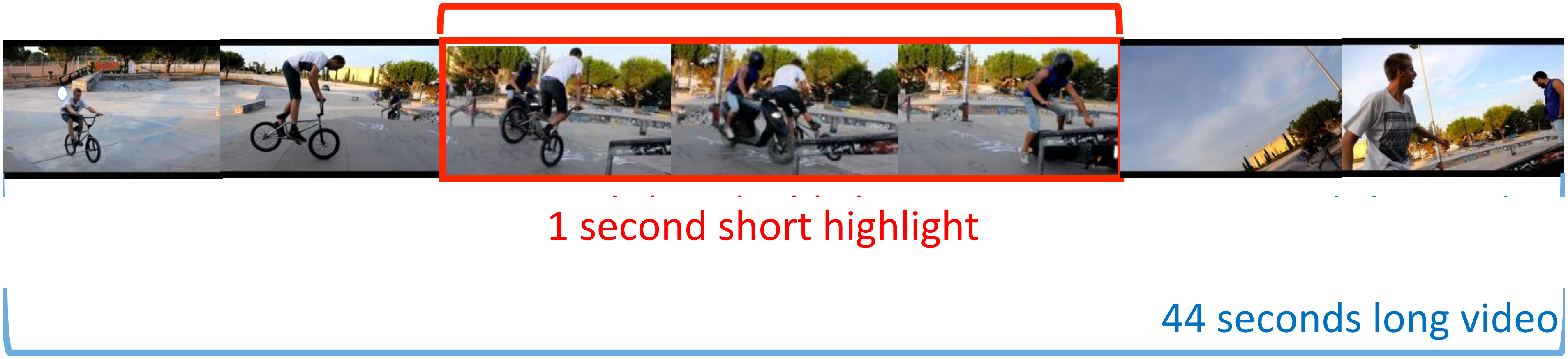
Pretty Good Title

Video Title Generation

Title vs. Caption

- Catchy
- Describing the most salient event (Highlight)

Title (most salient event): Bmx rider gets *hit by scooter* at park



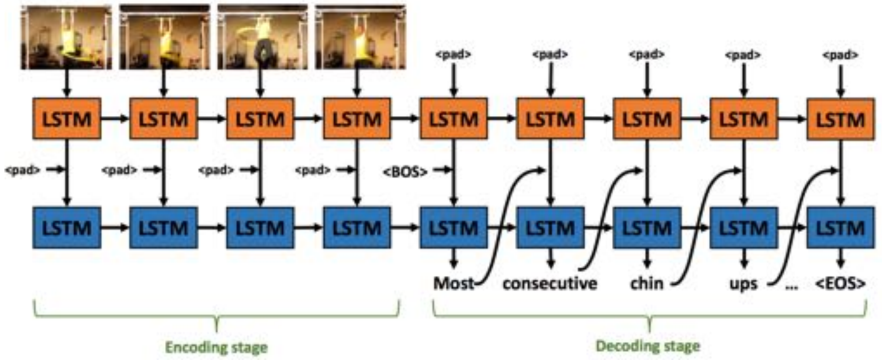
Captions: A man riding on bike. A man does a stunt on a bmx bike.

- Generic
- Describing a video as a whole

Related Work

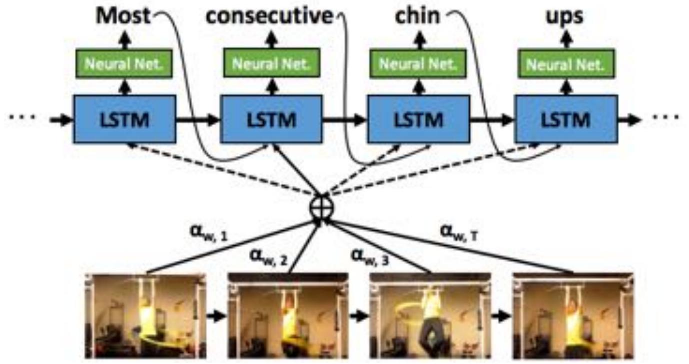
- Rohrbach et al. The long-short story of movie description. GCPR’15

- S2VT



Venugopalan et al. Sequence to Sequence – Video to Text, *ICCV'15*.

- SA

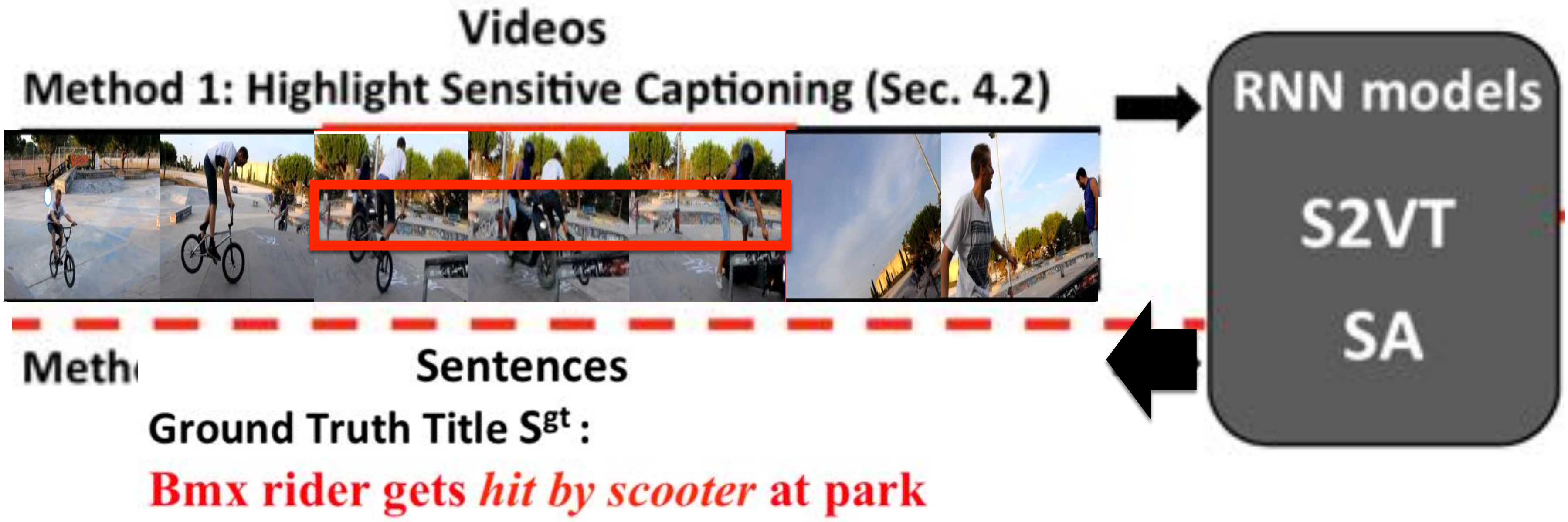


Yao et al. “Describing Videos by Exploiting Temporal Structure”, *ICCV'15*

- Pan et al. Jointly modeling embedding and translation to bridge video and language. CVPR’16
- Pan et al. Hierarchical recurrent neural encoder for video representation with application to captioning. CVPR’16
- Yu et al. Video paragraph captioning using hierarchical recurrent neural networks. CVPR’16

Video Title Generation

- Describing the most salient event (Highlight)
- Catchy

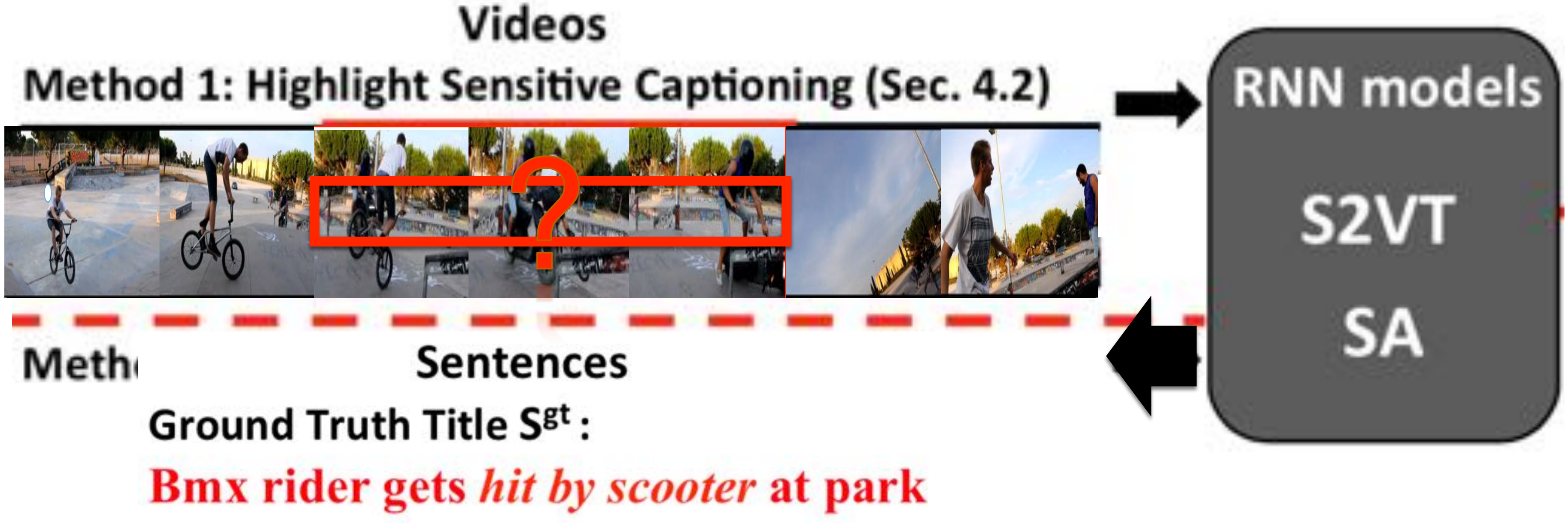


Ground Truth Title S^{gt} :

Bmx rider gets *hit by scooter* at park

Highlight Sensitive

- Describing the most salient event (highlight)
  - Unknown highlight location in training



Videos

Method 1: Highlight Sensitive Captioning (Sec. 4.2)

?

RNN models

S2VT

SA

Meth

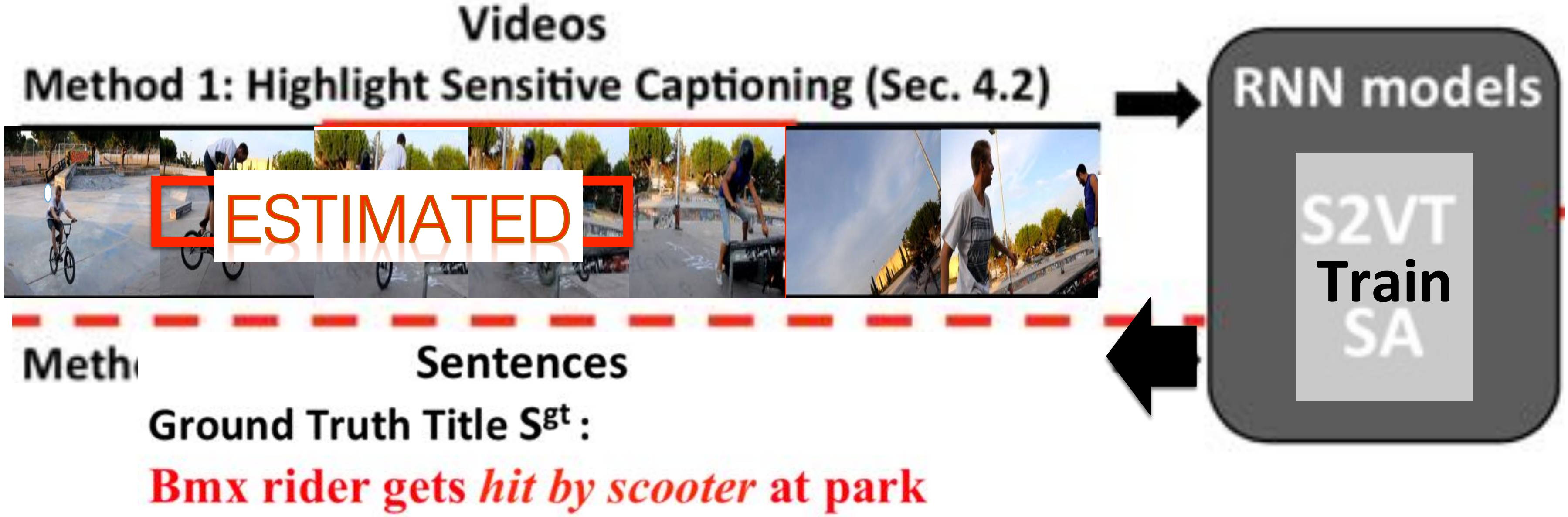
Sentences

Ground Truth Title S^{gt} :

Bmx rider gets *hit by scooter* at park

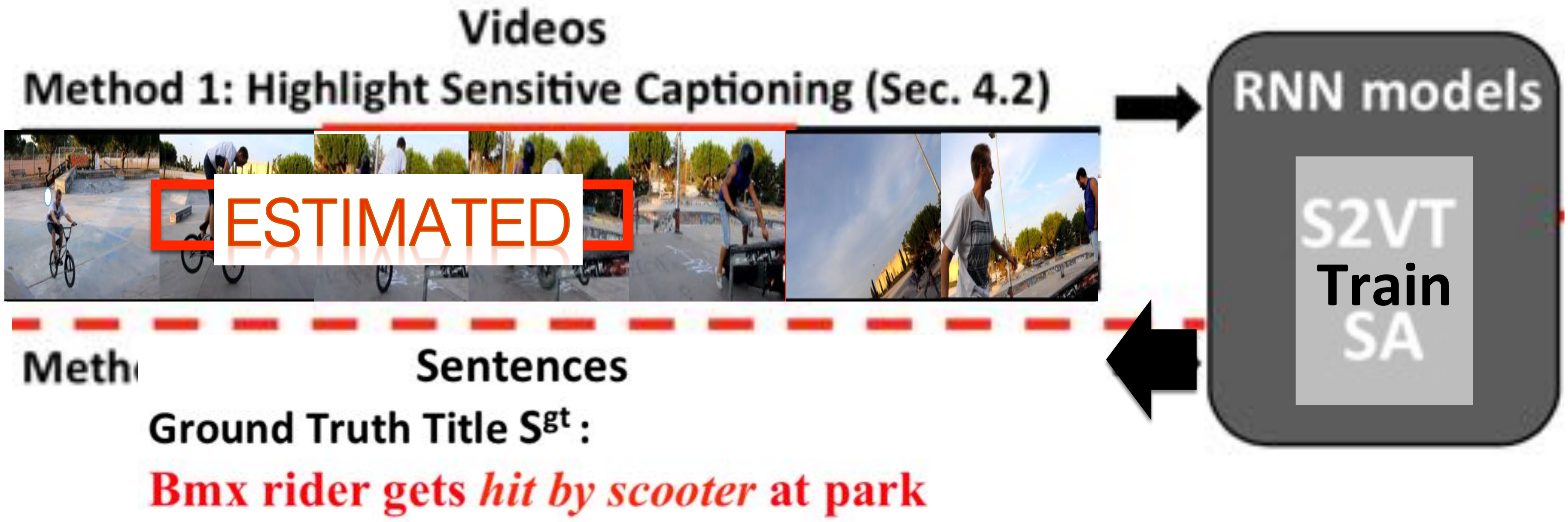
Highlight Sensitive

- Describing the most salient event (highlight)
  - Unknown highlight location in training


Videos
Method 1: Highlight Sensitive Captioning (Sec. 4.2)
ESTIMATED
RNN models
S2VT
Train
SA
Meth
Sentences
Ground Truth Title S^{gt} :
Bmx rider gets *hit by scooter* at park

Highlight Sensitive

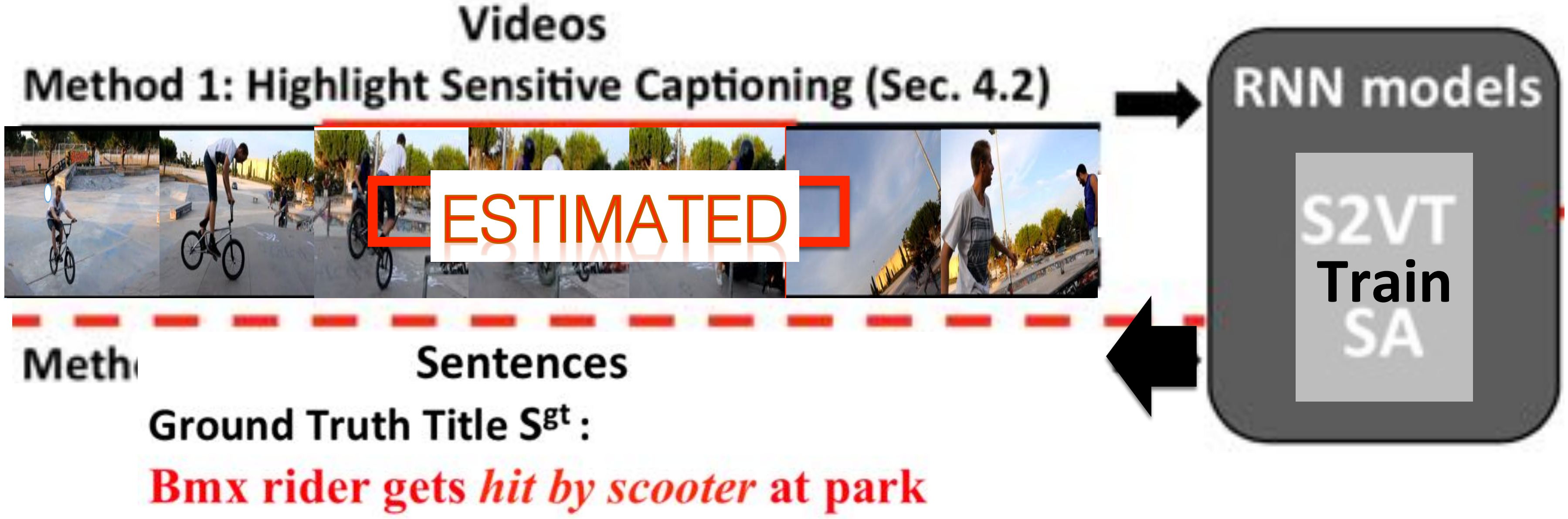
- Describing the most salient event (Highlight)
  - Unknown Highlight Location in Training


Videos
Method 1: Highlight Sensitive Captioning (Sec. 4.2)
ESTIMATED
RNN models
S2VT
Train
SA
Meth
Sentences
Ground Truth Title S^gt :
Bmx rider gets hit by scooter at park

VSLab

Highlight Sensitive

- Describing the most salient event (Highlight)
  - Unknown Highlight Location in Training


Videos
Method 1: Highlight Sensitive Captioning (Sec. 4.2)
ESTIMATED
RNN models
S2VT
Train
SA
Meth
Sentences
Ground Truth Title S^{gt} :
Bmx rider gets *hit by scooter* at park

VSLab

Video Title Generation

- Describing the most salient event (Highlight)
- Catchy (Diverse)

Sentences

Ground Truth Title S^{gt} :

Bmx rider gets *hit by scooter* at park

Web Sentence Augmentation

Mountain biker gets hit by car

Bmx rider performs bike transfer

Bmx rider pops wheelie and falls

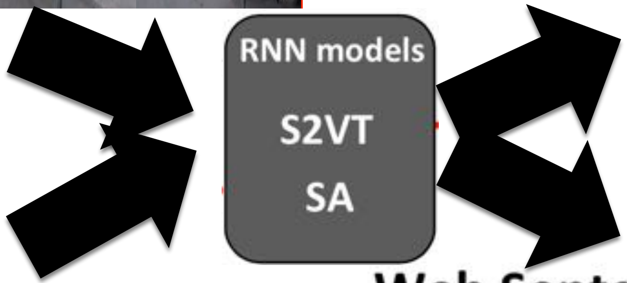
Sentence Augmentation

- Describing the most salient event (Highlight)
- Catchy (Diverse)

Sentences

Ground Truth Title S^{gt} :

Bmx rider gets *hit by scooter* at park



Web Sentence Augmentation

Mountain biker gets hit by car

Bmx rider performs bike transfer

Bmx rider pops wheelie and falls

Method 2: Sentence Augmentation (Sec. 4.3)

Dummy video observation v^D

Video Title in the Wild (VTW) Dataset

YouTube channels curating

- viral videos
- editor-verified video titles

Video Title in the Wild (VTW) Dataset

VSLab

Title: Kitten Falls off Dresser

Description: Just as this kitten started to get the nerve up to leap from the top of a dresser to the floor, it struggled with its balance and fell off.

Title: Hungry Baby Elephant Starts Tug of War with Tourist's Scarf

Description: This baby Indian elephant may look docile, but this tourist quickly learns otherwise — it's really a scarf-scarfing machine! While petting the elephant's trunk and sporadically turning to pose for her videographer husband, the woman suddenly finds herself in a fight for her scarf, now the subject of a tug-of-war match between herself and this hungry, hungry elephant.

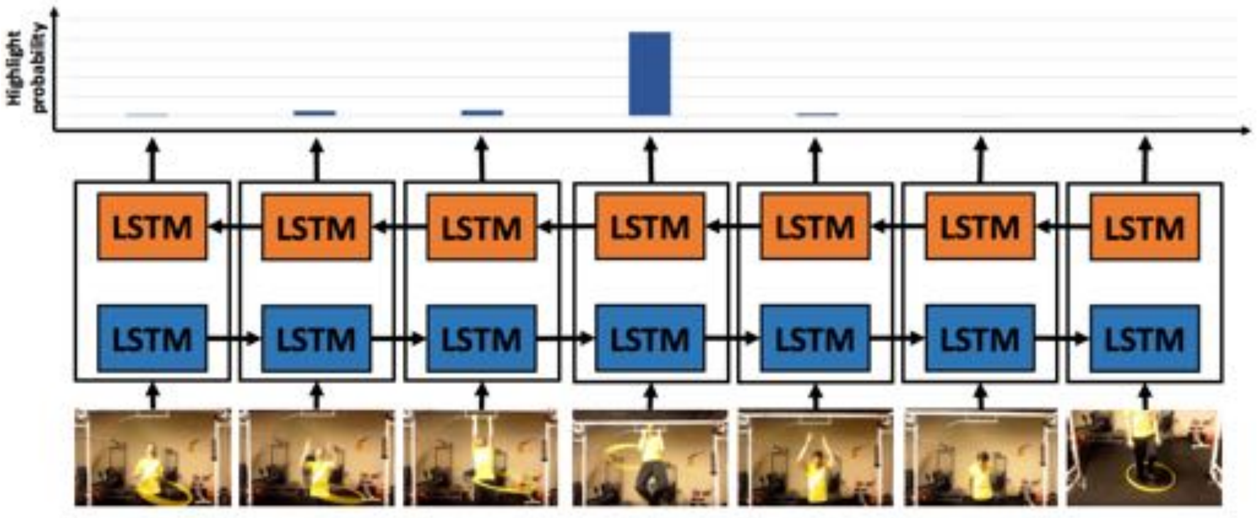
Video Title in the Wild (VTW) Dataset

VSLab

- **Videos**: 14100(training) - 2000(testing) - 2000(validation)
- **Titles**: 14100(training) - 2000(testing) - 2000(validation)

Details

- Initial weak highlight detector (1000 training videos)



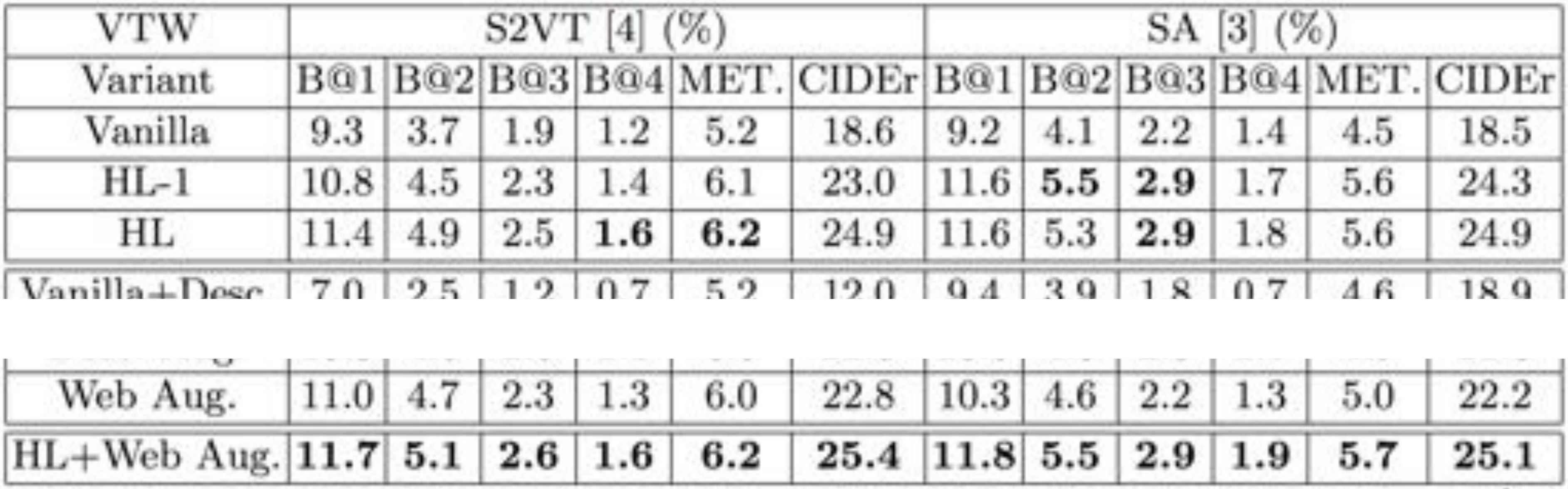
- Augment from **Web** (3546 sentences)

FAILARMY™

PRANK

Result on VTW

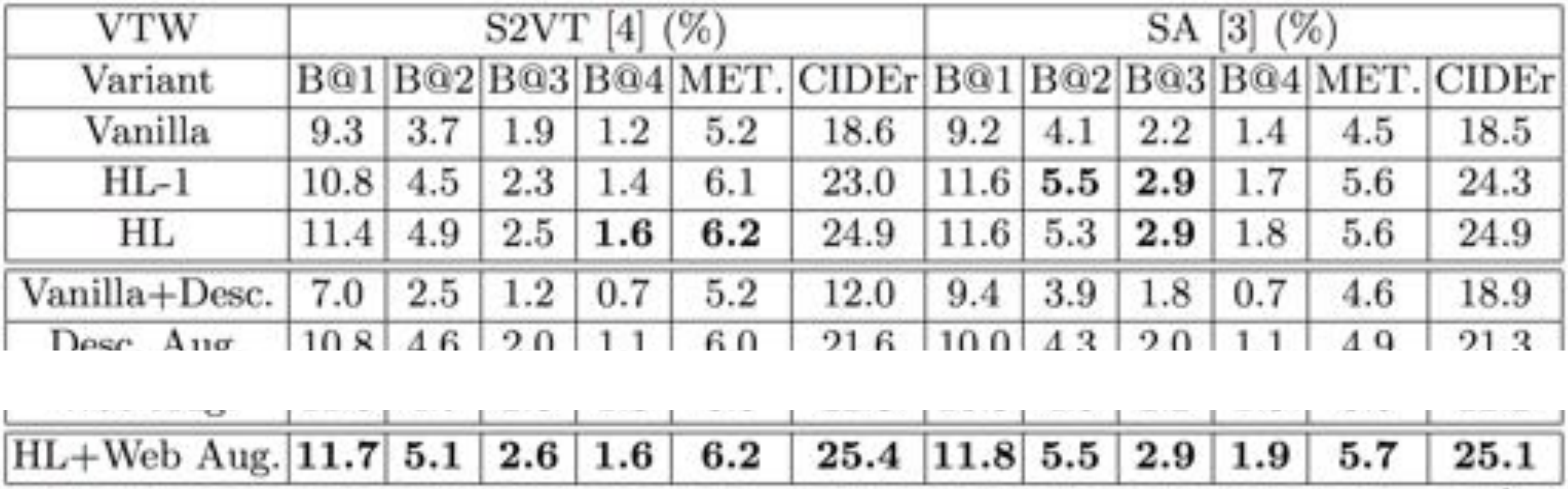
VSLab



VTW	S2VT [4] (%)						SA [3] (%)					
Variant	B@1	B@2	B@3	B@4	MET.	CIDEr	B@1	B@2	B@3	B@4	MET.	CIDEr
Vanilla	9.3	3.7	1.9	1.2	5.2	18.6	9.2	4.1	2.2	1.4	4.5	18.5
HL-1	10.8	4.5	2.3	1.4	6.1	23.0	11.6	**5.5**	**2.9**	1.7	5.6	24.3
HL	11.4	4.9	2.5	**1.6**	**6.2**	24.9	11.6	5.3	**2.9**	1.8	5.6	24.9
Vanilla+Desc.	7.0	2.5	1.2	0.7	5.2	12.0	9.4	3.9	1.8	0.7	4.6	18.9
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Web Aug.	11.0	4.7	2.3	1.3	6.0	22.8	10.3	4.6	2.2	1.3	5.0	22.2
HL+Web Aug.	**11.7**	**5.1**	**2.6**	**1.6**	**6.2**	**25.4**	**11.8**	**5.5**	**2.9**	**1.9**	**5.7**	**25.1**

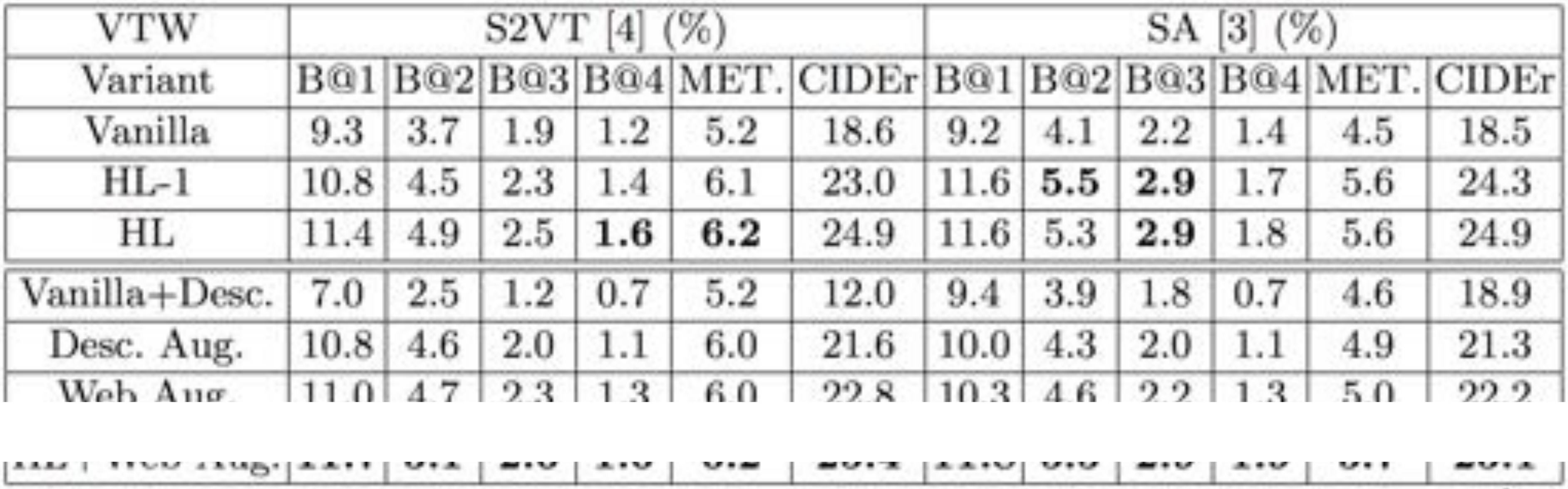
VSLab

Result on VTW



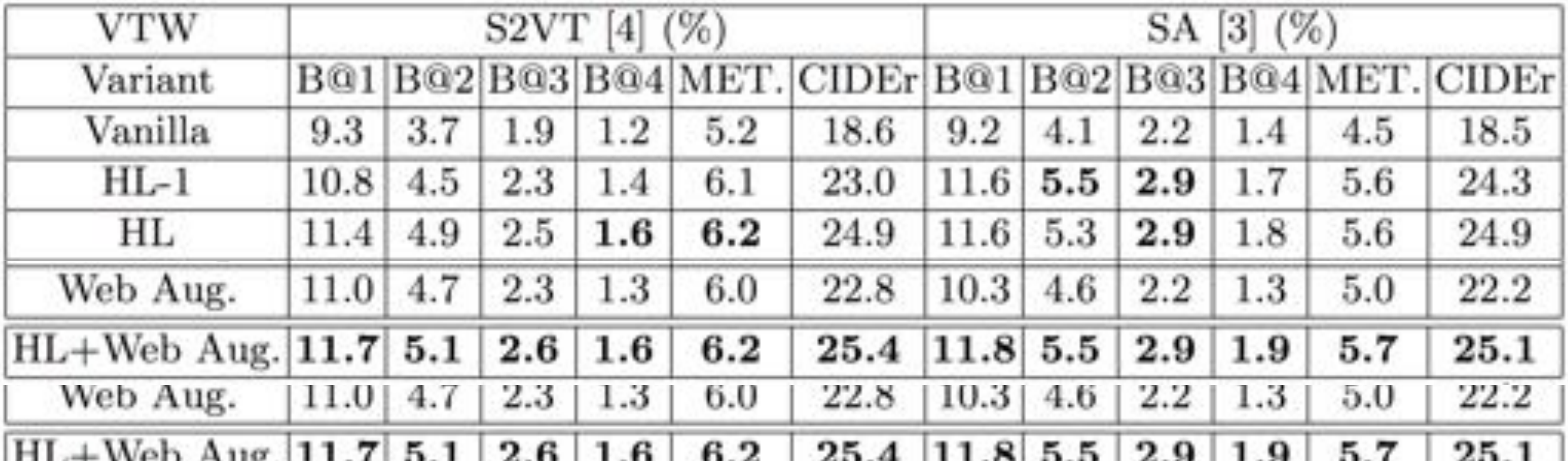
VTW	S2VT [4] (%)						SA [3] (%)					
Variant	B@1	B@2	B@3	B@4	MET.	CIDEr	B@1	B@2	B@3	B@4	MET.	CIDEr
Vanilla	9.3	3.7	1.9	1.2	5.2	18.6	9.2	4.1	2.2	1.4	4.5	18.5
HL-1	10.8	4.5	2.3	1.4	6.1	23.0	11.6	**5.5**	**2.9**	1.7	5.6	24.3
HL	11.4	4.9	2.5	**1.6**	**6.2**	24.9	11.6	5.3	**2.9**	1.8	5.6	24.9
Vanilla+Desc.	7.0	2.5	1.2	0.7	5.2	12.0	9.4	3.9	1.8	0.7	4.6	18.9
Desc. Aug	10.8	4.6	2.0	1.1	6.0	21.6	10.0	4.3	2.0	1.1	4.9	21.3
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
HL+Web Aug.	**11.7**	**5.1**	**2.6**	**1.6**	**6.2**	**25.4**	**11.8**	**5.5**	**2.9**	**1.9**	**5.7**	**25.1**

Result on VTW



VTW	S2VT [4] (%)						SA [3] (%)					
Variant	B@1	B@2	B@3	B@4	MET.	CIDEr	B@1	B@2	B@3	B@4	MET.	CIDEr
Vanilla	9.3	3.7	1.9	1.2	5.2	18.6	9.2	4.1	2.2	1.4	4.5	18.5
HL-1	10.8	4.5	2.3	1.4	6.1	23.0	11.6	**5.5**	**2.9**	1.7	5.6	24.3
HL	11.4	4.9	2.5	**1.6**	**6.2**	24.9	11.6	5.3	**2.9**	1.8	5.6	24.9
Vanilla+Desc.	7.0	2.5	1.2	0.7	5.2	12.0	9.4	3.9	1.8	0.7	4.6	18.9
Desc. Aug.	10.8	4.6	2.0	1.1	6.0	21.6	10.0	4.3	2.0	1.1	4.9	21.3
Web Aug.	11.0	4.7	2.3	1.3	6.0	22.8	10.3	4.6	2.2	1.3	5.0	22.2
HL + Web Aug.	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Result on VTW



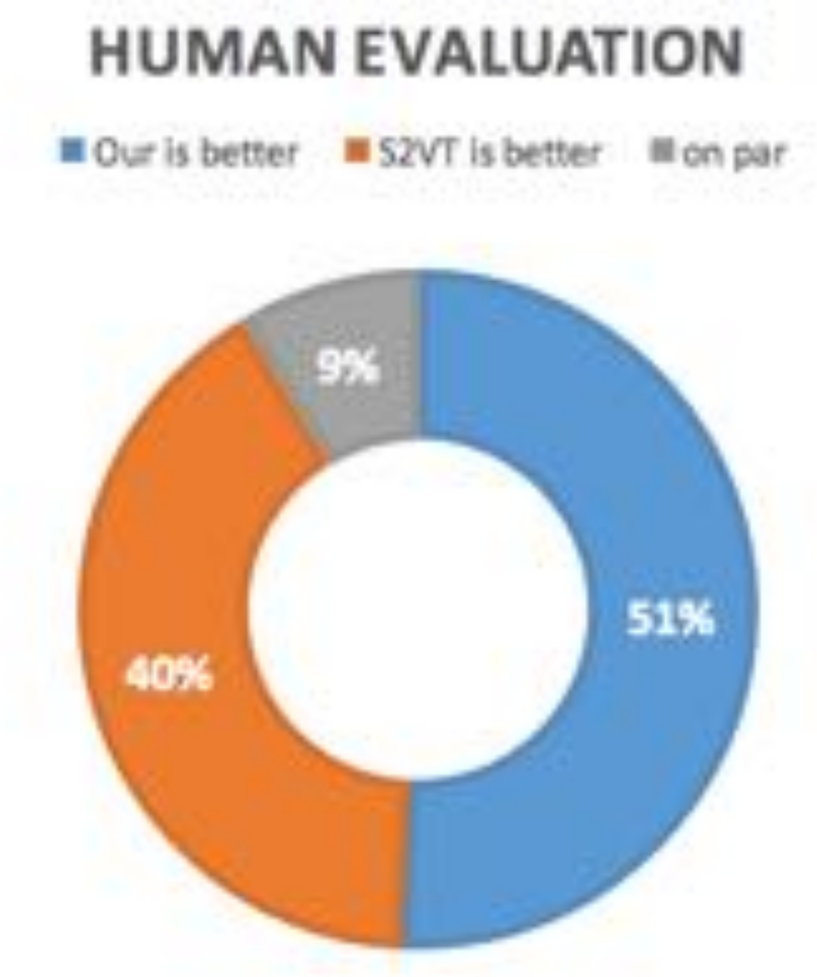
VTW	S2VT [4] (%)						SA [3] (%)					
Variant	B@1	B@2	B@3	B@4	MET.	CIDEr	B@1	B@2	B@3	B@4	MET.	CIDEr
Vanilla	9.3	3.7	1.9	1.2	5.2	18.6	9.2	4.1	2.2	1.4	4.5	18.5
HL-1	10.8	4.5	2.3	1.4	6.1	23.0	11.6	**5.5**	**2.9**	1.7	5.6	24.3
HL	11.4	4.9	2.5	**1.6**	**6.2**	24.9	11.6	5.3	**2.9**	1.8	5.6	24.9
Web Aug.	11.0	4.7	2.3	1.3	6.0	22.8	10.3	4.6	2.2	1.3	5.0	22.2
HL+Web Aug.	**11.7**	**5.1**	**2.6**	**1.6**	**6.2**	**25.4**	**11.8**	**5.5**	**2.9**	**1.9**	**5.7**	**25.1**

Web Aug. 11.0 4.7 2.3 1.3 6.0 22.8 10.3 4.6 2.2 1.3 5.0 22.2
HL+Web Aug. 11.7 5.1 2.6 1.6 6.2 25.4 11.8 5.5 2.9 1.9 5.7 25.1

VSLab

Result on VTW

VTW			(%)	
Variant	I	@4	MET.	CIDEr
Vanilla	(	.4	4.5	18.5
HL-1	1	.7	5.6	24.3
HL	1	.8	5.6	24.9
Web Aug.	1	.3	5.0	22.2
HL+Web Aug.	**1**	**.9**	**5.7**	**25.1**
Web Aug.	1	.3	5.0	22.2
HL+Web Aug.	**1**	**.9**	**5.7**	**25.1**



Result on M-VAD

- Augment from **MPII** dataset

AD: Abby gets in the basket.

Script: After a moment a frazzled Abby pops up in his place.

Mike leans over and sees how high they are.

Mike looks down to see – they are now fifteen feet above the ground.

Abby clasps her hands around his face and kisses him passionately.

For the first time in her life, she stops thinking and grabs Mike and kisses the hell out of him.

- METERO: S2VT+Aug 7.1% vs. S2VT 6.7%

Torabi, A., Pal, C.J., Larochelle, H., Courville, A.C.: Using descriptive video services to create a large data source for video annotation research. In: arXiv:1503.01070'15
Rohrbach, A., Rohrbach, M., Tandon, N., Schiele, B.: A dataset for movie description. In: CVPR'15

Title Generation for User Generated Videos

ECCV 2016

VSLab Microsoft Research 微软亚洲研究院 科 技 部 Ministry of Science and Technology NOVATEK Panasonic

Thanks!