POET: Product Oriented Video Captioner for E-Commerce

Team:Group 46

Asmita Chotani, Apurva Gupta, Chetan Chaku, Priya Nayak, Rutuja Oza

Table of Contents

01

Motivation

02

Innovation

03

Methodology

04

Datasets and Evaluation

05

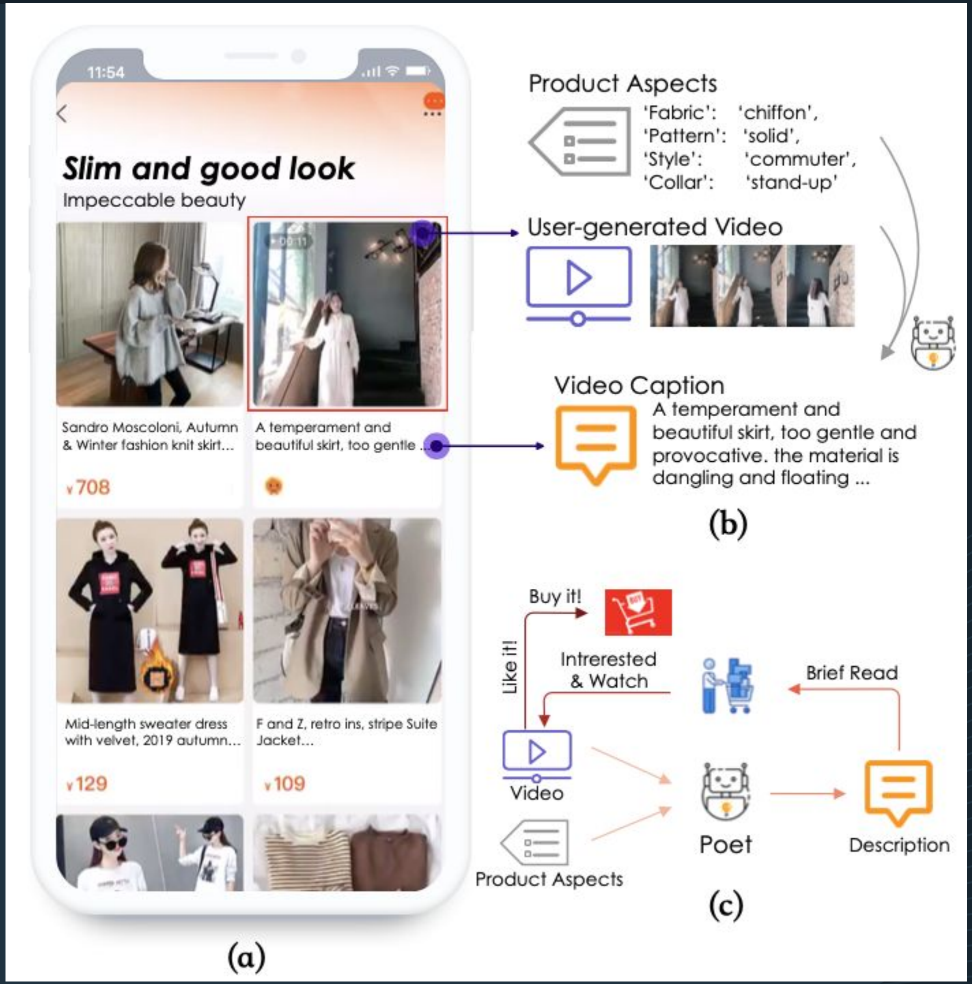
Performance Analysis

06

Conclusion

Motivation

1. Online Promotion of product by **User-generated Videos**
2. Showcase product features using captions
3. Time-consuming and expensive to create captions manually
4. Inaccurate or irrelevant captions by machine effects sale and customer satisfaction
5. POET aims to use **visual cues** from video and **product aspects** generating relevant captions


11:54
Slim and good look
Impeccable beauty
Sandro Moscoloni, Autumn & Winter fashion knit skirt...
A temperament and beautiful skirt, too gentle ...
¥708
Mid-length sweater dress with velvet, 2019 autumn...
F and Z, retro ins, stripe Suite Jacket...
¥129
¥109
(a)
Product Aspects
'Fabric': 'chiffon',
'Pattern': 'solid',
'Style': 'commuter',
'Collar': 'stand-up'
User-generated Video
Video Caption
A temperament and beautiful skirt, too gentle and provocative. the material is dangling and floating ...
(b)
Buy it!
Like it!
Intrrested & Watch
Brief Read
Video
Poet
Description
Product Aspects
(c)

Innovation

POET improves upon the existing mechanisms in these 2 aspects:

Video to Text Generation:

- Traditional V2T generation methods focus on Seq2Seq modeling
- Do not provide fine-grained analysis of video for caption generation
- Neglect spatial interactions between region-region and region-background within frames
  - POET performs product-part characteristic recognition
  - POET uses spatial-temporal graph to model these interactions

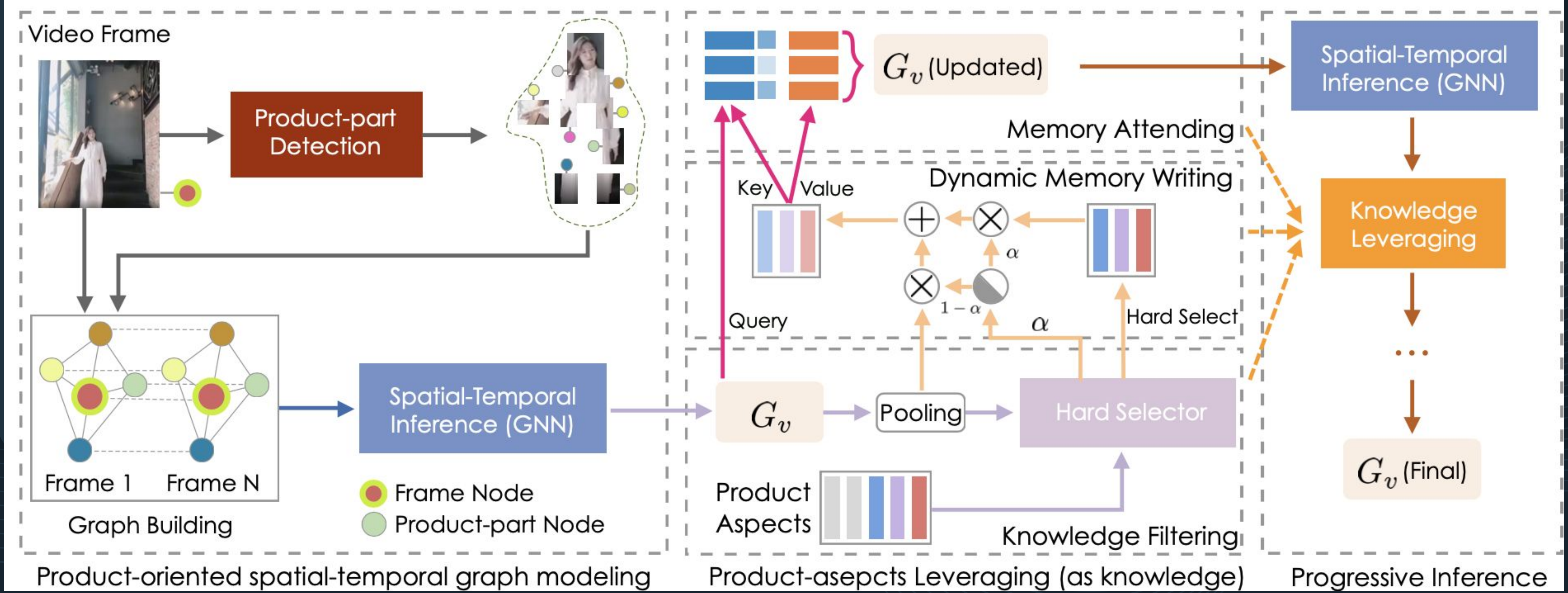
External Knowledge Leveraging:

- Traditional methods use off the shelf Knowledge graphs and document based approaches using pointer mechanisms to directly borrow entities during the decoding stage.
  - Poet performs knowledge leveraging in the product-oriented spatial-temporal inference stage using knowledge filtering and dynamic memory writing.

Methodology

Building blocks of POET include:

- Represent videos as spatial-temporal graphs
- Knowledge leveraging module
- Spatial-Temporal Inference Module
- Attentional RNN-based decoder



Datasets

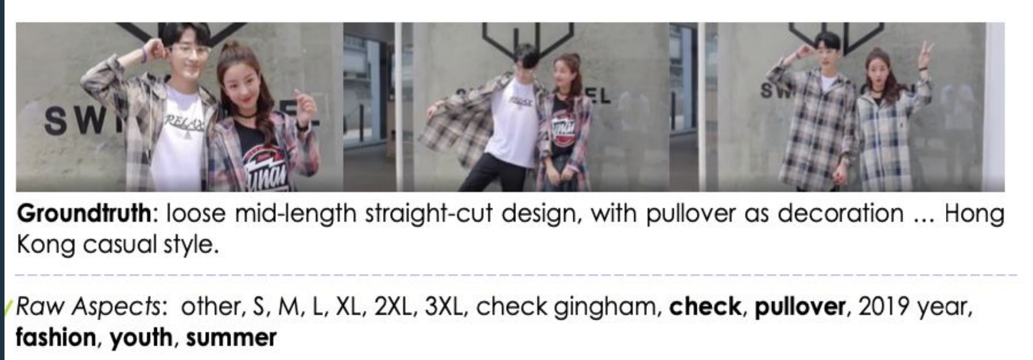
Collected Two Large Scale product oriented datasets from Mobile Taobao

Individual data points include video,product aspects and description of video(ground truth) triplets.

- Buyer Generated Fashion Video Dataset

  43,166 <video,description,aspect> triplets
- Fan Generated Fashion Video Dataset

  32,763 <video,description,aspect> triplets



Groundtruth: loose mid-length straight-cut design, with pullover as decoration ... Hong Kong casual style.

Raw Aspects: other, S, M, L, XL, 2XL, 3XL, check gingham, **check**, **pullover**, 2019 year, **fashion**, **youth**, **summer**

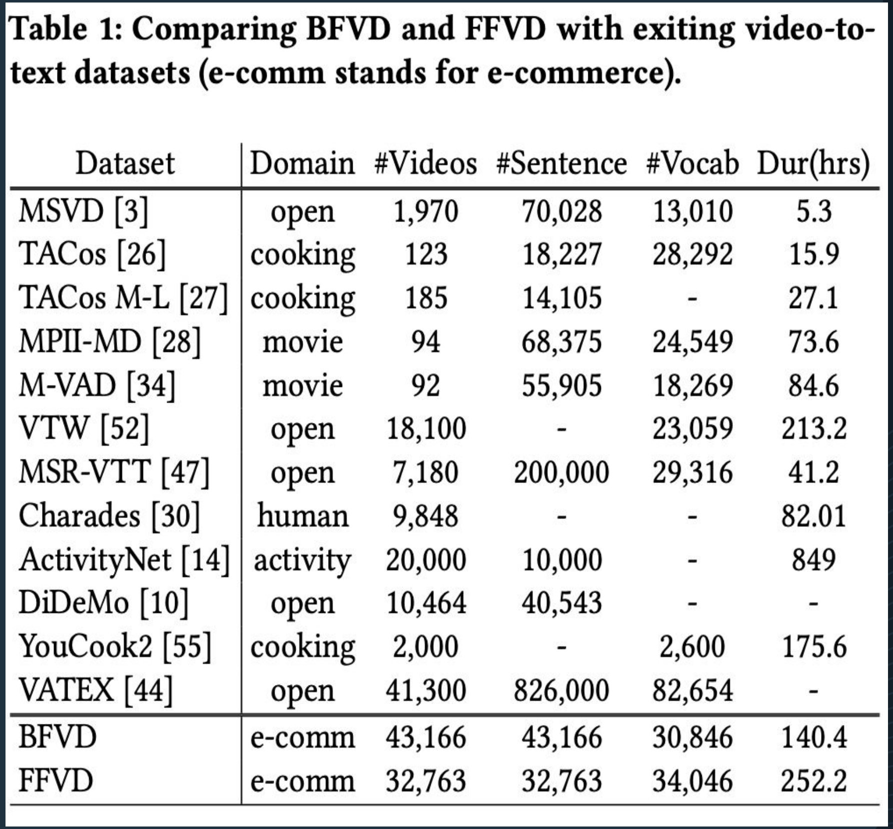


Table 1: Comparing BFVD and FFVD with exiting video-to-text datasets (e-comm stands for e-commerce).

Dataset	Domain	#Videos	#Sentence	#Vocab	Dur(hrs)
MSVD [3]	open	1,970	70,028	13,010	5.3
TACos [26]	cooking	123	18,227	28,292	15.9
TACos M-L [27]	cooking	185	14,105	-	27.1
MPII-MD [28]	movie	94	68,375	24,549	73.6
M-VAD [34]	movie	92	55,905	18,269	84.6
VTW [52]	open	18,100	-	23,059	213.2
MSR-VTT [47]	open	7,180	200,000	29,316	41.2
Charades [30]	human	9,848	-	-	82.01
ActivityNet [14]	activity	20,000	10,000	-	849
DiDeMo [10]	open	10,464	40,543	-	-
YouCook2 [55]	cooking	2,000	-	2,600	175.6
VATEX [44]	open	41,300	826,000	82,654	-
BFVD	e-comm	43,166	43,166	30,846	140.4
FFVD	e-comm	32,763	32,763	34,046	252.2

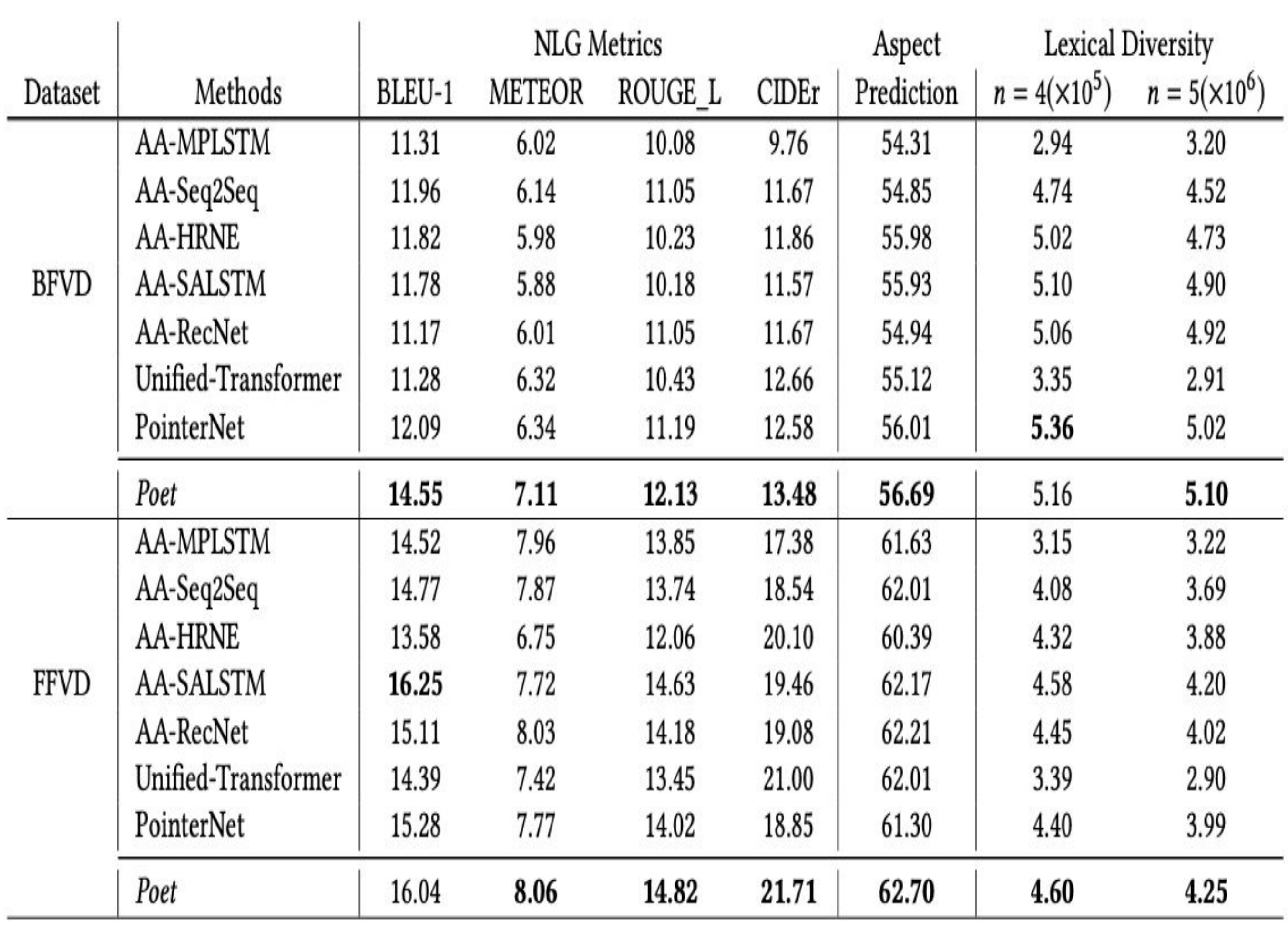
Evaluation

Evaluation Metrics:

- **Natural language generation Metrics:**
  - BLEU,METEOR,ROUGE,CIDEr for generation fluency
- **Aspect Prediction :**
  - to evaluate how product aspect knowledge is leveraged
- **Lexical Diversity:**
  - evaluate generation diversity through N-grams

Comparison Baselines:

- Re-implement 7 baselines adding separate encoders for product aspect modeling



Dataset	Methods	NLG Metrics				Aspect	Lexical Diversity	
		BLEU-1	METEOR	ROUGE_L	CIDEr	Prediction	$n = 4(\times 10^5)$	$n = 5(\times 10^6)$
BFVD	AA-MPLSTM	11.31	6.02	10.08	9.76	54.31	2.94	3.20
	AA-Seq2Seq	11.96	6.14	11.05	11.67	54.85	4.74	4.52
	AA-HRNE	11.82	5.98	10.23	11.86	55.98	5.02	4.73
	AA-SALSTM	11.78	5.88	10.18	11.57	55.93	5.10	4.90
	AA-RecNet	11.17	6.01	11.05	11.67	54.94	5.06	4.92
	Unified-Transformer	11.28	6.32	10.43	12.66	55.12	3.35	2.91
	PointerNet	12.09	6.34	11.19	12.58	56.01	**5.36**	5.02
	Poet	**14.55**	**7.11**	**12.13**	**13.48**	**56.69**	5.16	**5.10**
FFVD	AA-MPLSTM	14.52	7.96	13.85	17.38	61.63	3.15	3.22
	AA-Seq2Seq	14.77	7.87	13.74	18.54	62.01	4.08	3.69
	AA-HRNE	13.58	6.75	12.06	20.10	60.39	4.32	3.88
	AA-SALSTM	**16.25**	7.72	14.63	19.46	62.17	4.58	4.20
	AA-RecNet	15.11	8.03	14.18	19.08	62.21	4.45	4.02
	Unified-Transformer	14.39	7.42	13.45	21.00	62.01	3.39	2.90
	PointerNet	15.28	7.77	14.02	18.85	61.30	4.40	3.99
	Poet	16.04	**8.06**	**14.82**	**21.71**	**62.70**	**4.60**	**4.25**

Performance Analysis

Performance Analysis was performed on our 2 video-datasets, i.e., BFVD and FFVD on the following perspective:

- Human Evaluation:
  a. Perform the human evaluation as captions are highly diverse and creative.
  b. Transformer based model(AA-Transformer) and RNN based model(AA-RecNet) were compared with Poet
  c. Fluency, Diversity, and Overall Quality are 3 characteristics used for comparison
- Parameter Analysis
  a. Ablation Studies
     i. It is used to check the effectiveness of the proposed modules within the Poet.
     ii. Adding pointer mechanism and removing knowledge leverage in dataset clearly drop the performance.

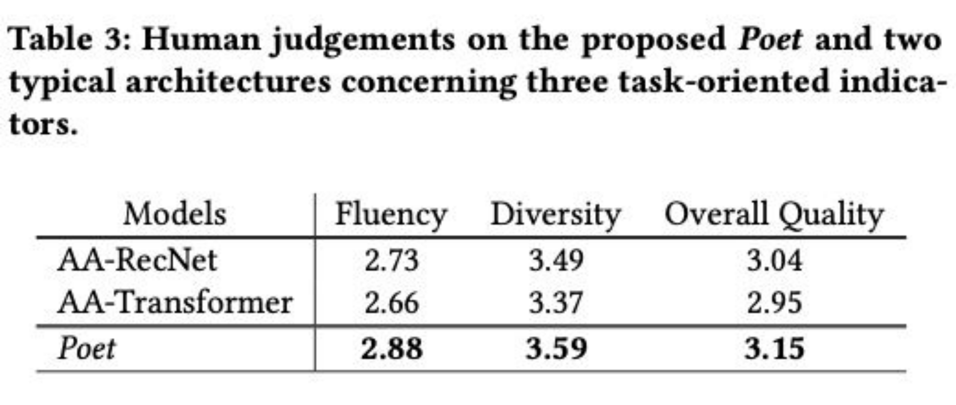


Table 3: Human judgements on the proposed *Poet* and two typical architectures concerning three task-oriented indicators.

Models	Fluency	Diversity	Overall Quality
AA-RecNet	2.73	3.49	3.04
AA-Transformer	2.66	3.37	2.95
Poet	**2.88**	**3.59**	**3.15**

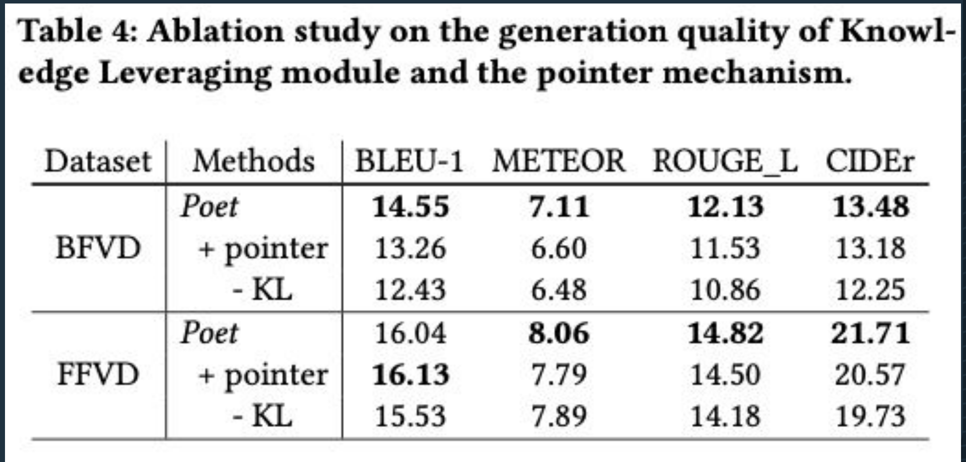


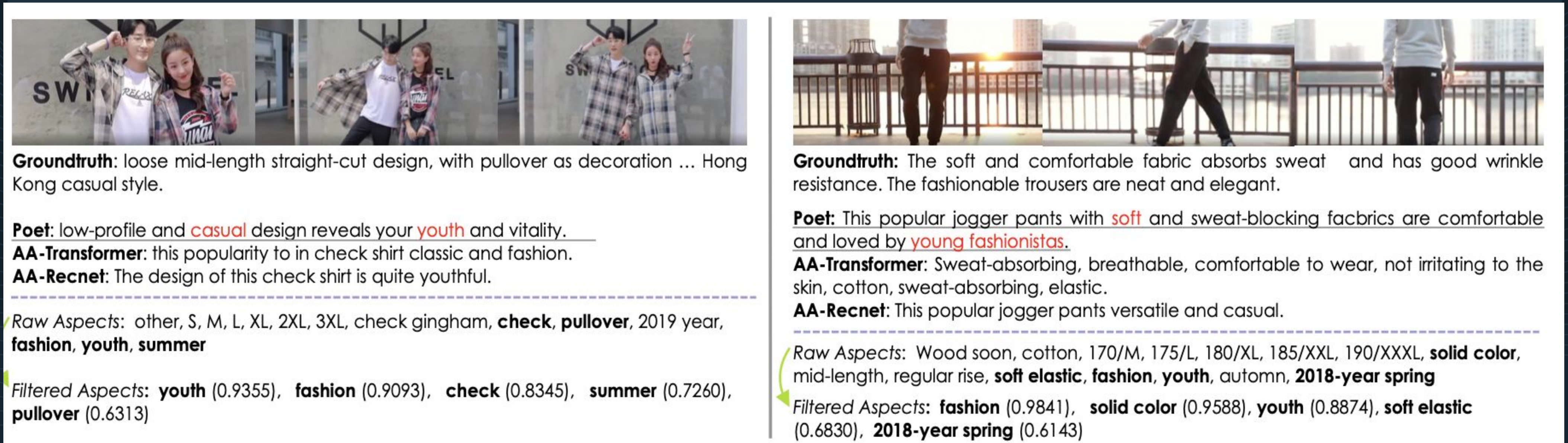
Table 4: Ablation study on the generation quality of Knowledge Leveraging module and the pointer mechanism.

Dataset	Methods	BLEU-1	METEOR	ROUGE_L	CIDEr
BFVD	*Poet*	**14.55**	**7.11**	**12.13**	**13.48**
	+ pointer	13.26	6.60	11.53	13.18
	- KL	12.43	6.48	10.86	12.25
FFVD	*Poet*	16.04	**8.06**	**14.82**	**21.71**
	+ pointer	**16.13**	7.79	14.50	20.57
	- KL	15.53	7.89	14.18	19.73

Conclusion

The authors built the framework POET performing knowledge-enhanced spatial-temporal inference on product-oriented video graphs. Advantages of POET include:

- Automates Video Generation Process
- Generates fluent,complete and relatively diversified sentences
- Capable of generating creative and accurate words from given aspects
- Capable of making the decision of which aspects would attract customers.



Groundtruth: loose mid-length straight-cut design, with pullover as decoration ... Hong Kong casual style.

Poet: low-profile and casual design reveals your youth and vitality.
AA-Transformer: this popularity to in check shirt classic and fashion.
AA-Recnet: The design of this check shirt is quite youthful.

Raw Aspects: other, S, M, L, XL, 2XL, 3XL, check gingham, **check**, **pullover**, 2019 year, **fashion**, **youth**, **summer**

Filtered Aspects: **youth** (0.9355), **fashion** (0.9093), **check** (0.8345), **summer** (0.7260), **pullover** (0.6313)

Groundtruth: The soft and comfortable fabric absorbs sweat and has good wrinkle resistance. The fashionable trousers are neat and elegant.

Poet: This popular jogger pants with soft and sweat-blocking facbrics are comfortable and loved by young fashionistas.
AA-Transformer: Sweat-absorbing, breathable, comfortable to wear, not irritating to the skin, cotton, sweat-absorbing, elastic.
AA-Recnet: This popular jogger pants versatile and casual.

Raw Aspects: Wood soon, cotton, 170/M, 175/L, 180/XL, 185/XXL, 190/XXXL, **solid color**, mid-length, regular rise, **soft elastic**, **fashion**, **youth**, automn, **2018-year spring**

Filtered Aspects: **fashion** (0.9841), **solid color** (0.9588), **youth** (0.8874), **soft elastic** (0.6830), **2018-year spring** (0.6143)