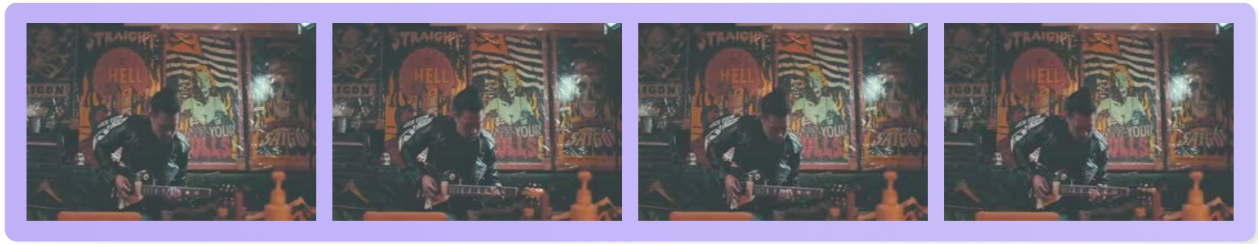
RUCAIM3-Tencent at TRECVID 2022:

Video-to-Text Description

Zihao Yue, Yuqi Liu, Liang Zhang, Linli Yao, Qin Jin
Renmin University of China
Dec.7, 2022

Video-to-Text Description (VTT)

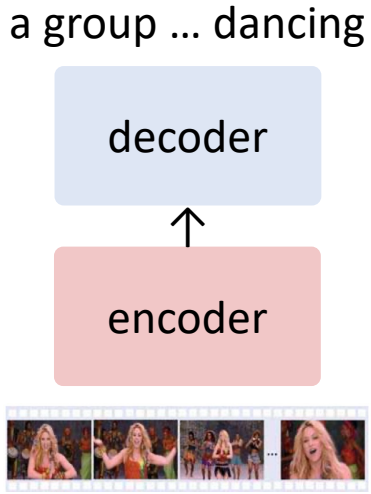
- automatically generate a single-sentence description in natural language for a given video.



- *An Asian man playing an electronic guitar in an indoor setting.*

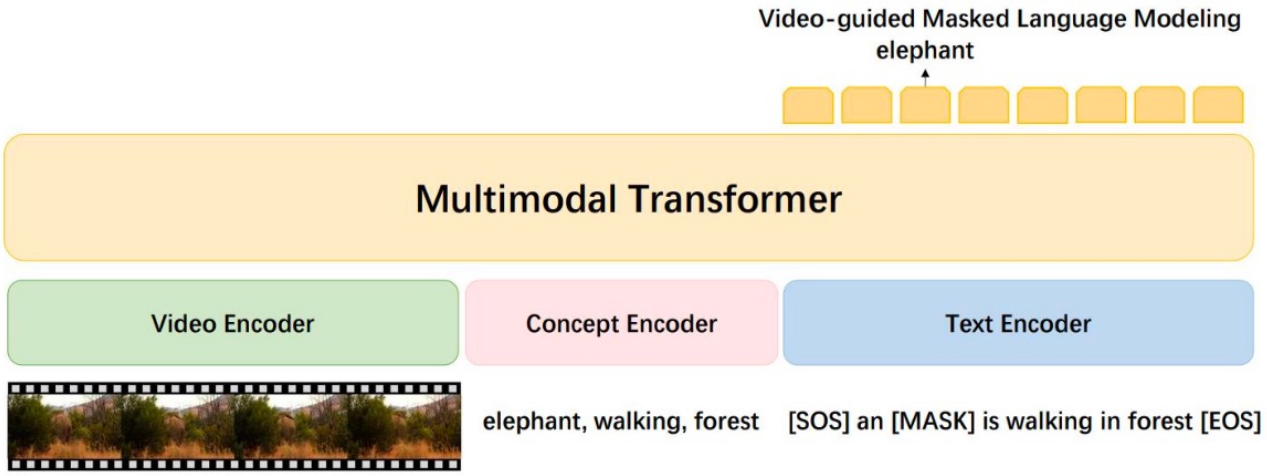
Dominant Approach: encoder-decoder framework

- encoder encodes videos into visual representations
- decoder generates captions conditioned on the encoder output



Introduction

our last year's solution (Zhang, et al.): Concept-Enhanced Pre-training-based Model (CE-PTM)



- Concept Encoder: encode concept (from an off-the-shelf concept extractor) representations
- Bert-like pre-training task (Video-guided Masked Language Modeling, VMLM)
- next-token prediction for fine-tuning (Modified VMLM)
- best CIDEr score: 36.0, ranking 1st

Introduction

Vision-Language Pre-training (VLP) Models

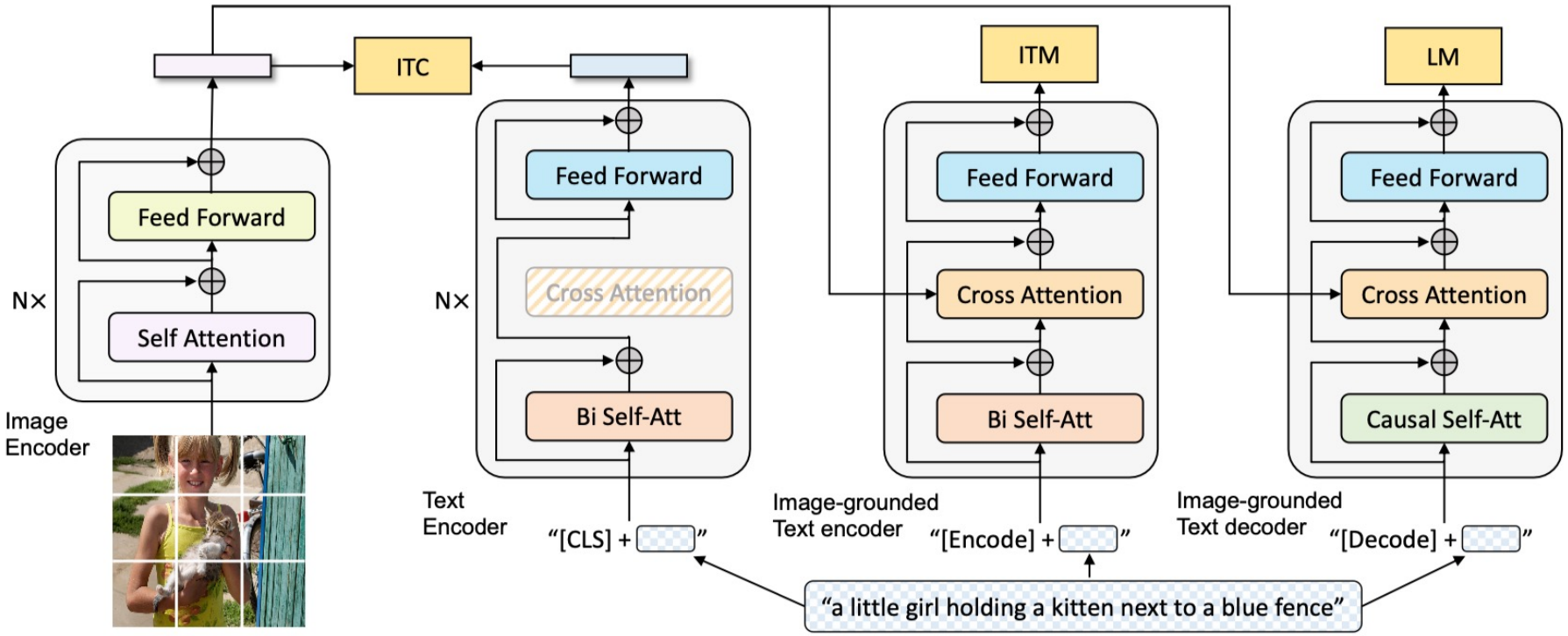
- UniVL (Luo, et al., 2020), Oscar (Li, et al., 2020), CLIP (Radford, et al., 2021) …
- learn effective representations from large-scale image-text data

Leveraging VLP Models for VTT

image-text models or video-text models?

- the amount of video-text data is not as large as that of image-text data
- we therefore consider to leverage image-text pre-trained models for videos tasks
- our choice is BLIP (Li et al., 2022)

Introduction - BLIP



- Bootstrapping Language-Image Pre-training
- multi-task: image-text contrastive, image-text matching, language modeling
- pre-trained on a bootstrapped dataset with 129M images and paired captions

Methodology - Overview

BLIP4video

our model structure

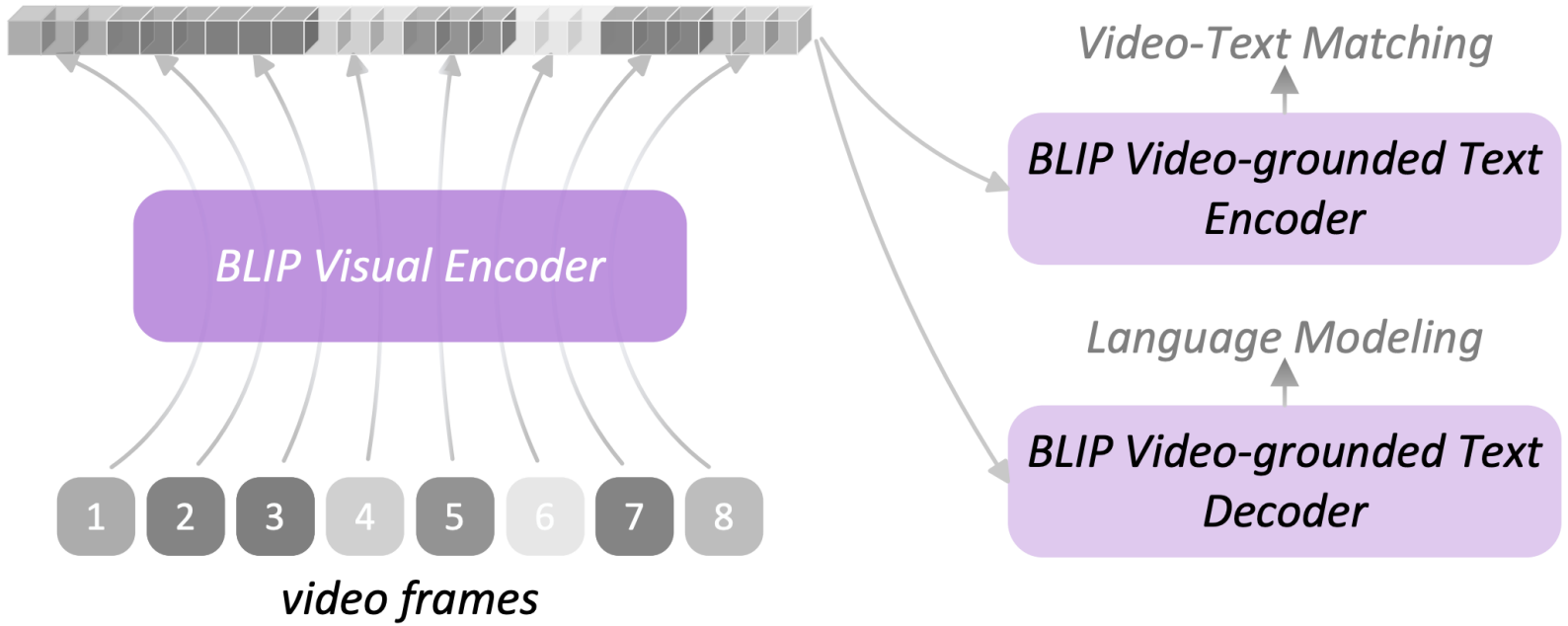
Data Augmentation

solving the problem of insufficient fine-tuning data

Candidates Re-ranking

best candidates selection

BLIP4video: structurally identical to BLIP, but supports the input of frame sequence

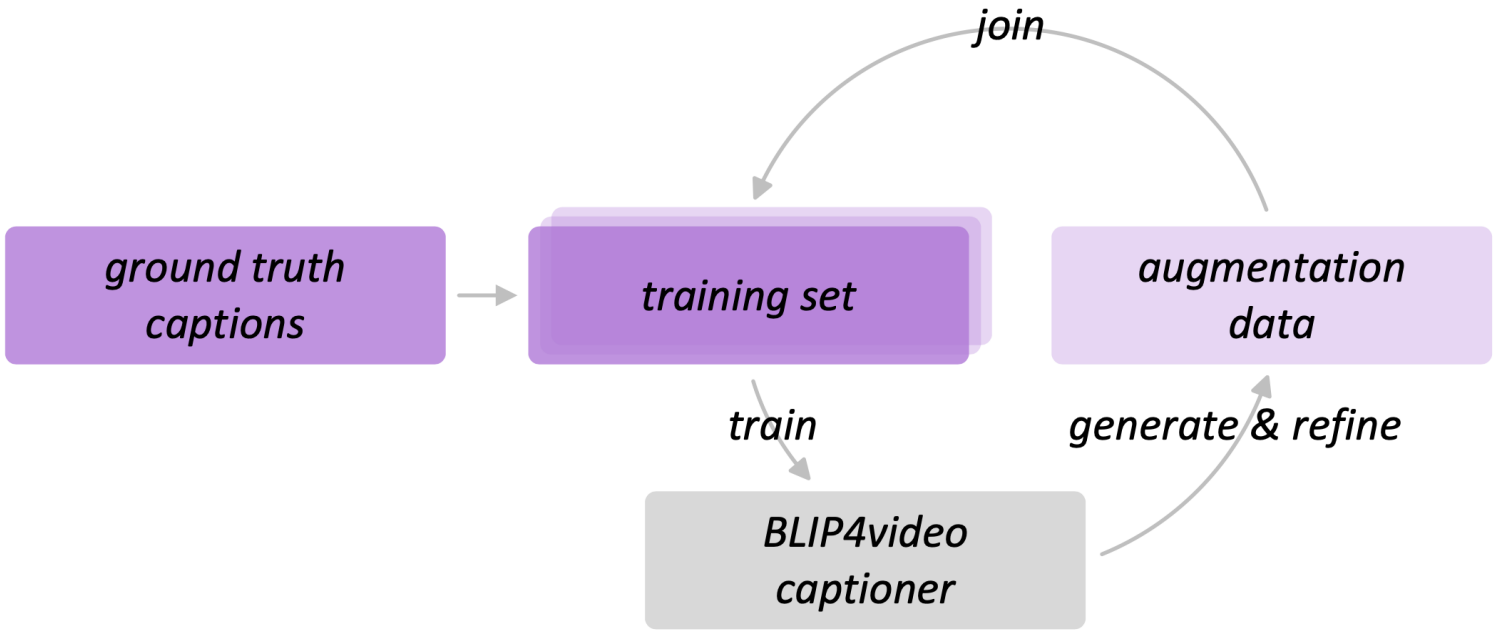


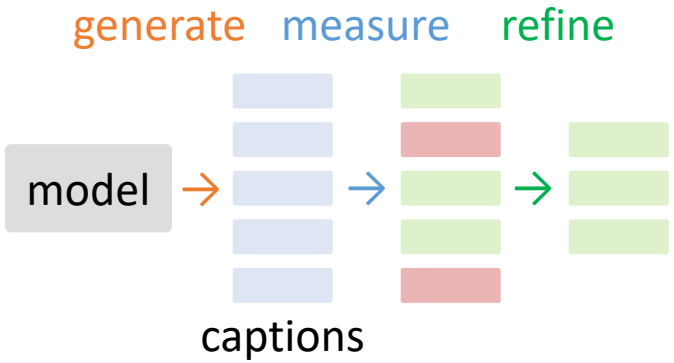
video-grounded text decoder: next-token prediction for caption generation

video-grounded text encoder: calculates a matching score between video and text

Methodology - Data Augmentation

pseudo-label-based data augmentation





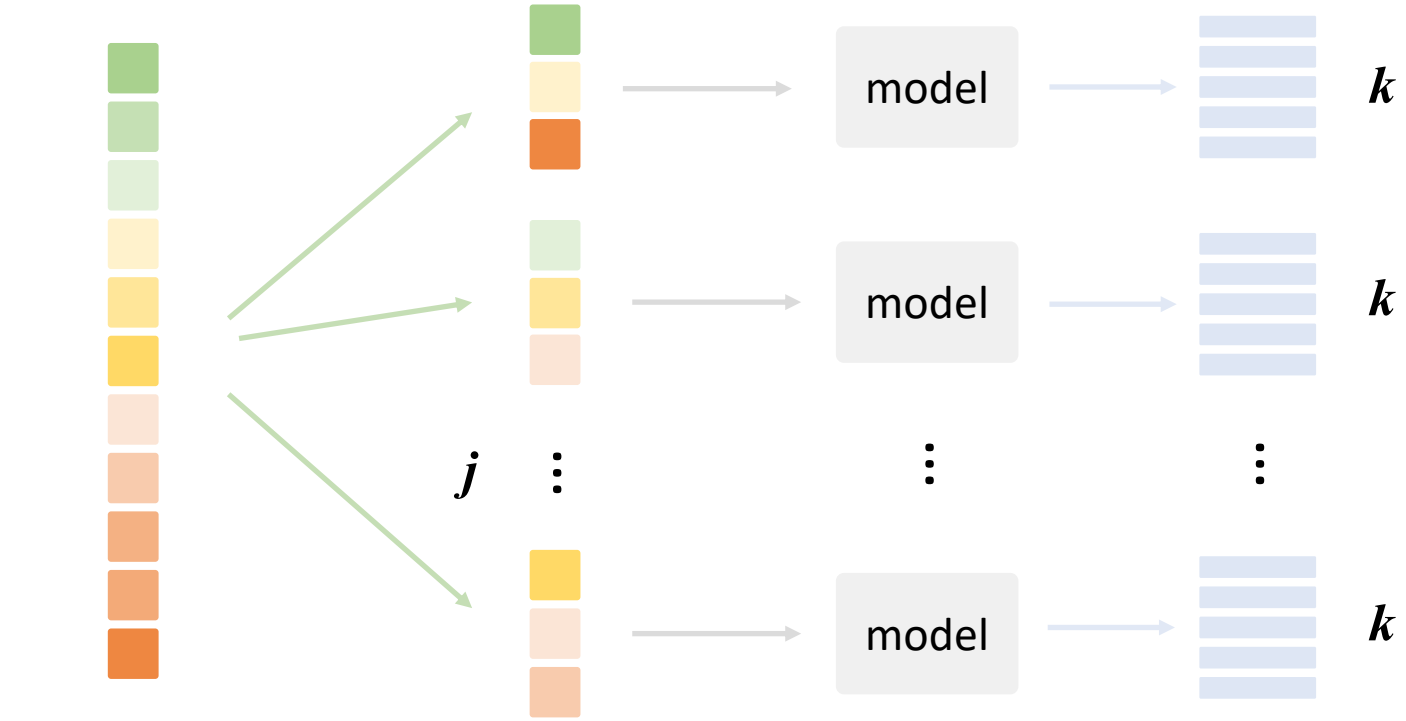
generation: 5 captions per video via beam search decoding

refinement: calculating CIDEr scores and filtering with a threshold

Methodology - Candidates Re-ranking

$j \times k$ candidates per video

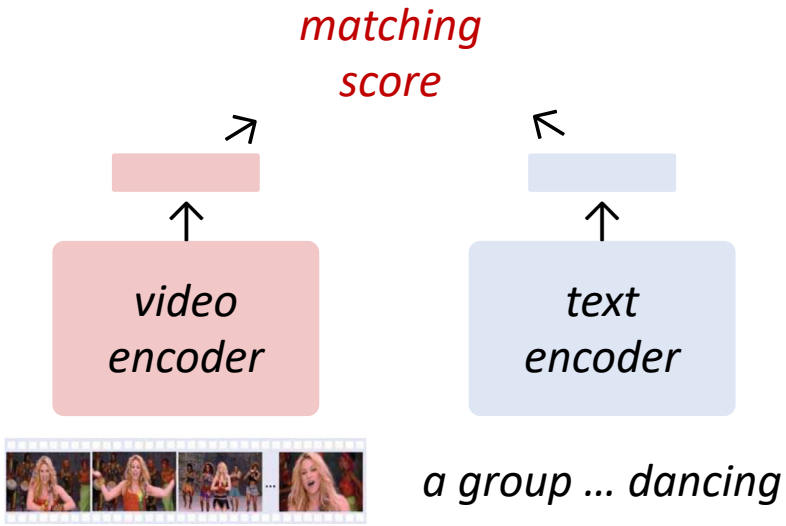
- j inferences $I_{1,2,\ldots,j}$ with different randomly selected frames as input for each video
- k sentences $S_{i1,i2,\ldots,ik}$ for each inference



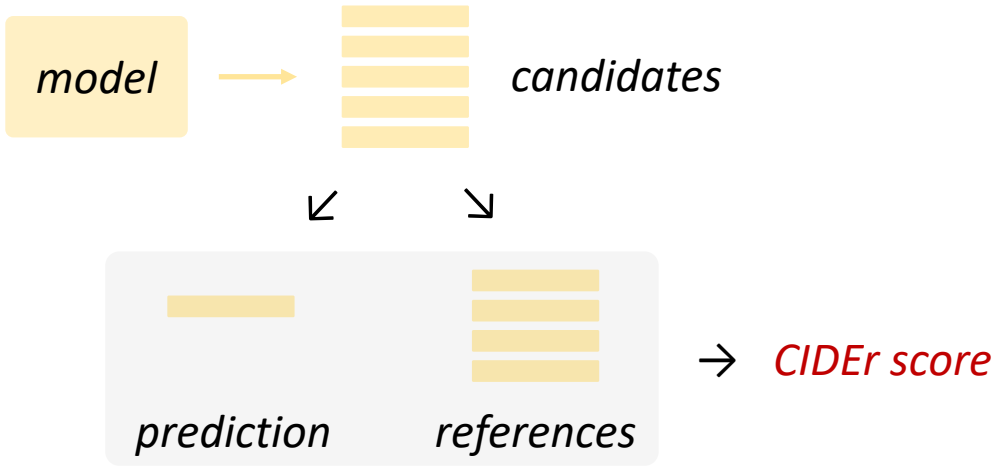
Methodology - Candidates Re-ranking

two measures

- Cross-modal Matching (CMM)
- Mutual Similarity Evaluation (MSE)



cross-modal matching

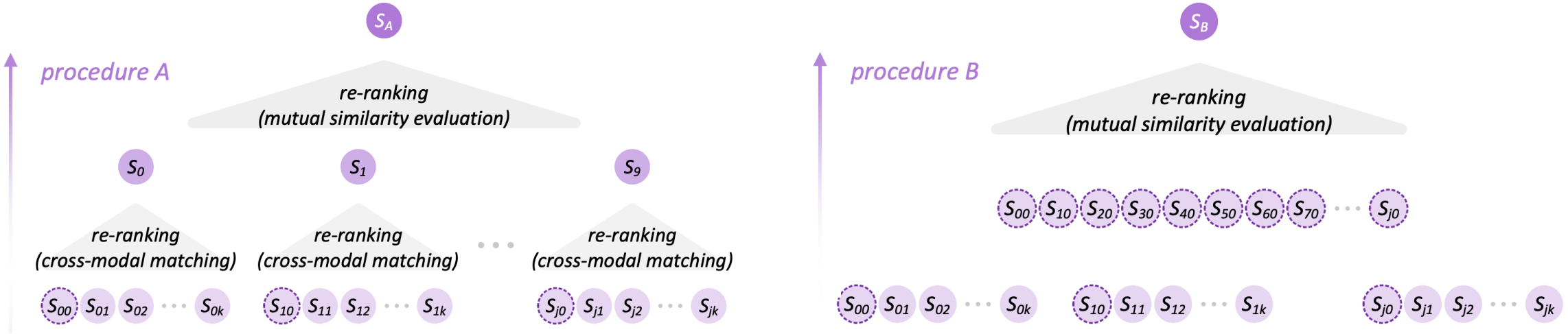


mutual similarity evaluation

Methodology - Candidates Re-ranking

two procedures

- CMM intra-inference & MSE inter-inference <CMM, MSE>
- top-beam intra-inference and MSE inter-inference <top-beam, MSE>



Experiment - Implementation Details

video input details

- 8 frames of 224×224 per video
- TSN sampling during training
- uniform sampling during inference

TSN sampling (Wang, et al., 2016) divides the video equally into *k* segments and select one frame from each randomly

training details (3-stage training)

- stage-1 (if adopted): training with *Extended datasets*
- stage-2: training with *VTT data* (and *augmentation data*)
- stage-3: SCST with *VTT data*

SCST (Rennie, et al., 2017) Self-critical Sequence Training we implement SCST as in VinVL (Zhang, et al., 2021)

Experiment - Implementation Details

2 rounds of data aumentation

- **1st round**
  - learning *Aug-Model-1* from *extended datasets*, to generate *Aug-1*
  - *Aug-Model-1*'s validation performance: 51.8 (CIDEr)
  - *Aug-1*: 17,939 captions adopted (~48% of *VTT16-20*)
- **2nd round**
  - learning *Aug-Model-2* from *VTT datasets* and *Aug-1*, to generate *Aug-2*
  - *Aug-Model-2*'s validation performance: 52.8 (CIDEr)
  - *Aug-2*: 18,784 captions adopted (~50% of *VTT16-20*)

nearly doubling the training data scale

Experiment - Implementation Details

2 final version models

- *Final-Model-1*: fine-tuned with VTT data and augmentation data
- *Final-Model-2*: fine-tuned with VTT data, augmentation data and validation set

4 runs

- *run1, run2*: generated by *Final-Model-1*; re-ranked by procedure A and B, respectively
- *run3, run4*: generated by *Final-model-2*; re-ranked by procedure A and B, respectively

Run Name	Model					Re-ranking procedure
	Name	Dataset			CIDEr	
		Training stage-2	Training stage-3	Validation		
run1 run2	Final-Model-1	VTT16-20, Aug1-2$_{C>50}$ (74,158)	VTT16-20 (37,435)	VTT21 (8,385)	53.9	A B
run3 run4	Final-Model-2	VTT16-21, Aug1-2$_{C>60}$ (74,845)	VTT16-21 (45,820)	-	-	A B

Experiment - Main Results

Run Name	Validation Performance				Submission Scores				
	CIDEr	BLEU@4	METEOR	ROUGE_L	CIDEr	BLEU@4	METEOR	SPICE	STS
run1	**54.9**	28.8	**22.4**	46.4	59.4	**13.5**	41.2	18.2	53.0
run2	54.1	**29.5**	22.2	**46.6**	57.5	13.2	40.9	18.0	52.8
run3	-	-	-	-	**60.2**	13.3	**41.5**	**18.4**	**53.4**
run4	-	-	-	-	59.2	**13.5**	41.4	18.3	53.0

- our VLP-based solutions set new performance records
- using both training and validation data, run3&4 outperform run1&2
- both re-ranking procedures excel in different evaluation metrics respectively
  - <top-beam, MSE> performs better on BLEU@4
  - <CMM, MSE> better on others

Experiment - Ablation Study

Row	Data Augmentation		Training Strategy	Re-ranking		Perform.
	Aug-1	Aug-2	SCST	A	B	CIDEr
1						48.2
2	✓					52.8
3	✓	✓				53.6
4	✓	✓	✓			53.9
5	✓	✓	✓	✓		**54.9**
6	✓	✓	✓		✓	54.1

Ablation Study of key components

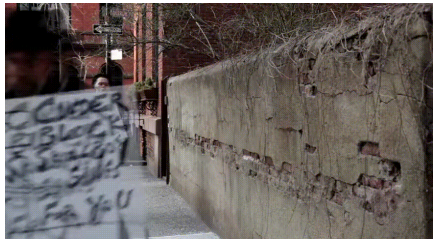
- data augmentation contributes a lot
- SCST is helpful
- re-ranking is helpful
  - <CMM, MSE> works better on CIDEr

Experiment - Cases

a man is playing a guitar in a room with posters on the wall

a black bird with a red beak is standing on the rocks near the ocean on a sunny day

a silhouette of a woman is shown in a profile against a white background



a man in a suit and hat is holding a sign in front of a stone wall on a city street

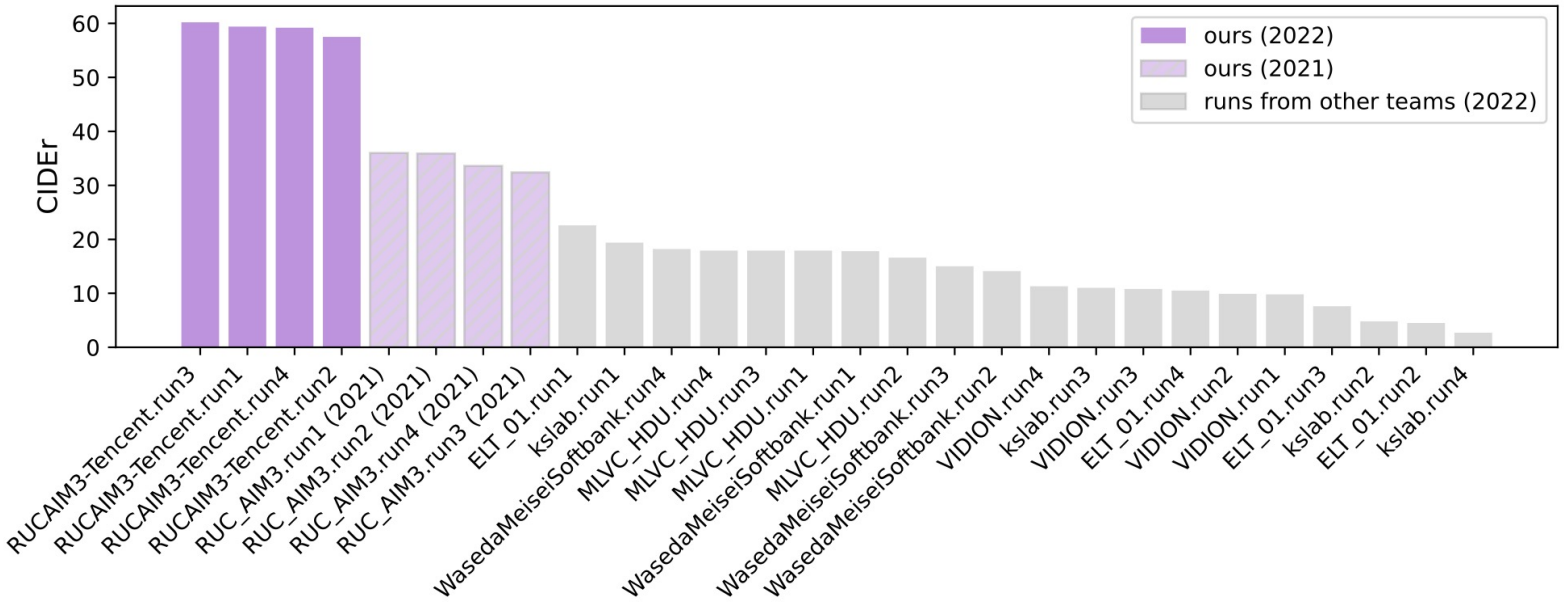
Conclusion

RUCAIM3-Tencent's solutions for VTT

- a strong baseline by fine-tuning a VLP model on the VTT task
- effective data augmentation and candidates re-ranking strategies

ranks 1st in all evaluation metrics (BLEU, METEOR, CIDER, SPICE, and STS)

best CIDEr score: 60.2 (67.2% higher than last year's best result)



Is our model good enough?

BLIP4video

- each frame is encoded separately to a final visual representation
- lacking inter-frame dynamics encoding

future works

- better video representation learning
  - extending the visual understanding capabilities of VLP models from images to videos
  - video-text pre-training

Is our benchmark challenging enough?

lack of *videoness*

- many videos in the VTT data are informatively static
- image captioning systems are easily competent for videos
- a more challenging benchmark requires cases with more *videoness*

- *An Asian man playing an electronic guitar in an indoor setting.*
- *An Asian guitarist plays his guitar inside a bar with colorful posters on the wall.*
- *An Asian man in a black jacket playing a guitar indoors.*
- *A man wearing a black jacket is playing a guitar indoors.*
- *Young man with black hair wearing black leather jacket plays electric guitar inside a dark room surrounded by posters.*

Limitations - Evaluation Metrics

Are our evaluation metrics reasonable enough?

Group	GT		Prediction	
	CIDEr	BLEU@4	CIDEr	BLEU@4
1	36.8	16.7	51.3	25.6
2	36.3	16.4	50.4	25.5
3	36.0	16.2	50.1	25.6
4	37.5	16.9	49.8	25.7
5	36.9	17.0	50.9	25.5
avg.	36.7	16.6	50.5	25.6

how GTs score on these metrics

- human experts score much less than models

existing metrics

- increasingly fail to measure the accuracy of generated descriptions
- other aspects (including fluency and diversity) are ignored
- representation learning-based metrics (CLIP score, et al.)? poor interpretability

future work

- exploring better evaluation metrics for the VTT task

Thanks!

Feel free to contact us:
yzihao@ruc.edu.cn
qjin@ruc.edu.cn