BUPT MCPRL

BUPT-MCPRL at TRECVID 2023 Video to Text Description

ZeLiang Ma , Shuai Jiang , Zhe Cui, Yanyun Zhao
Beijing University of Posts and Telecommunications
{mzl, js, cuizhe, zyy}@bupt.edu.cn

北京邮電大学
Beijing University of Posts and Telecommunications

CONTENTS

1.Challenge

2.Method

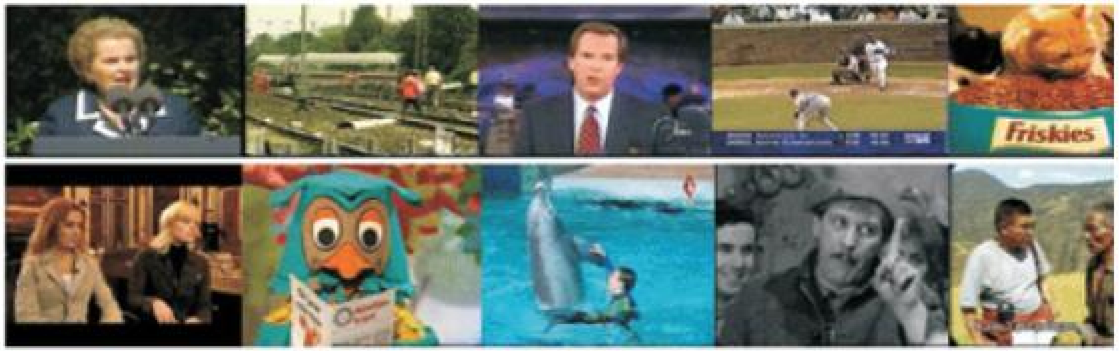
3.Results

4.Conclusion

北京郵電大學

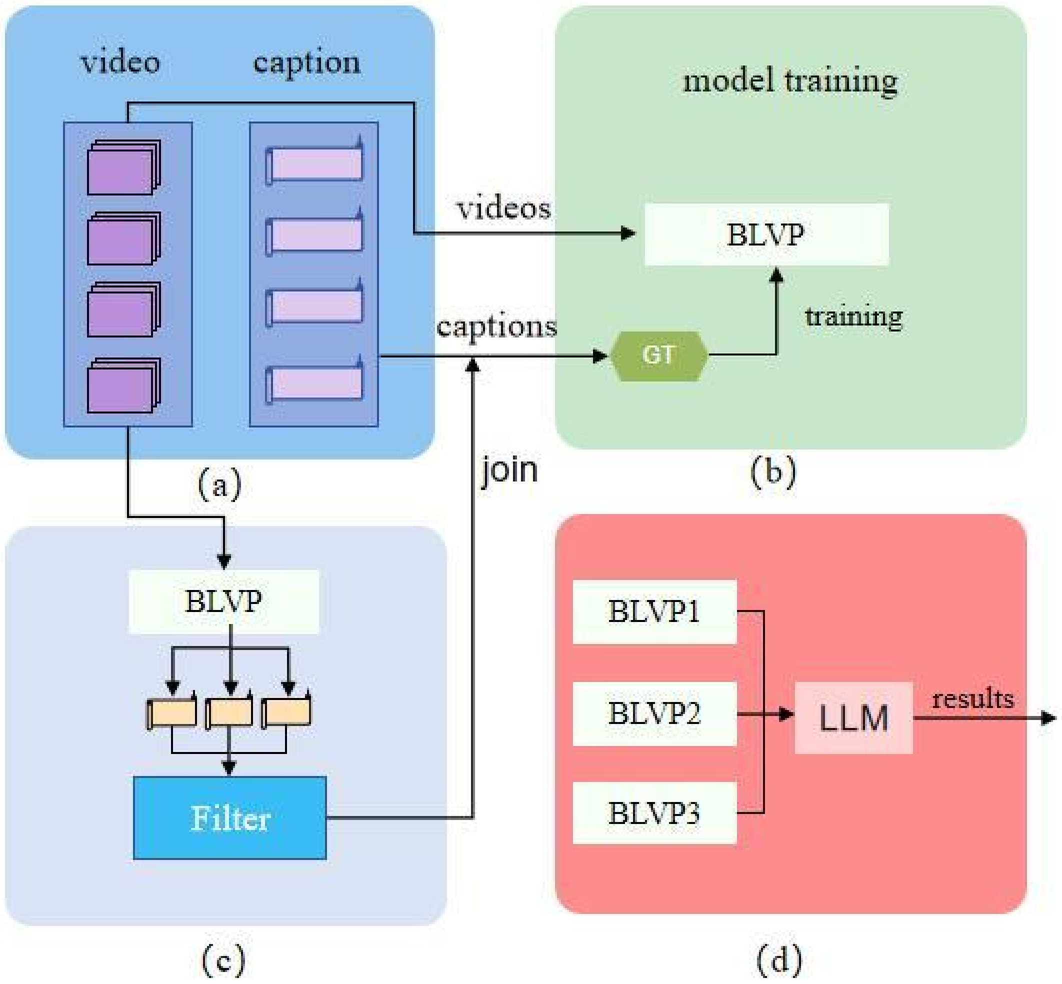
Beijing University of Posts and Telecommunications

01 Challenges



02 Method

- Overall workflow of our system



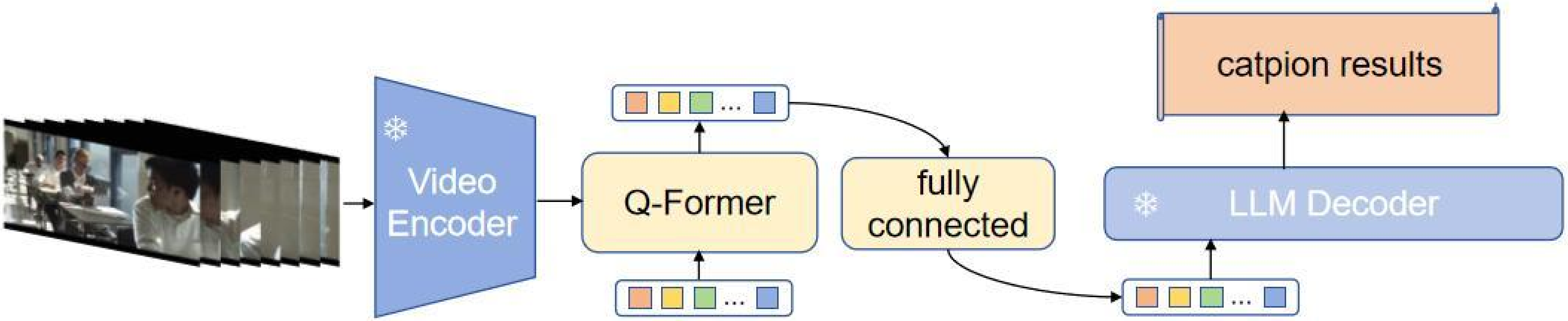
(a) the raw data

(b) the process of model training

(c) Process of New Data Generation and Data Filtering

(d)the multi-model fusion of LLM

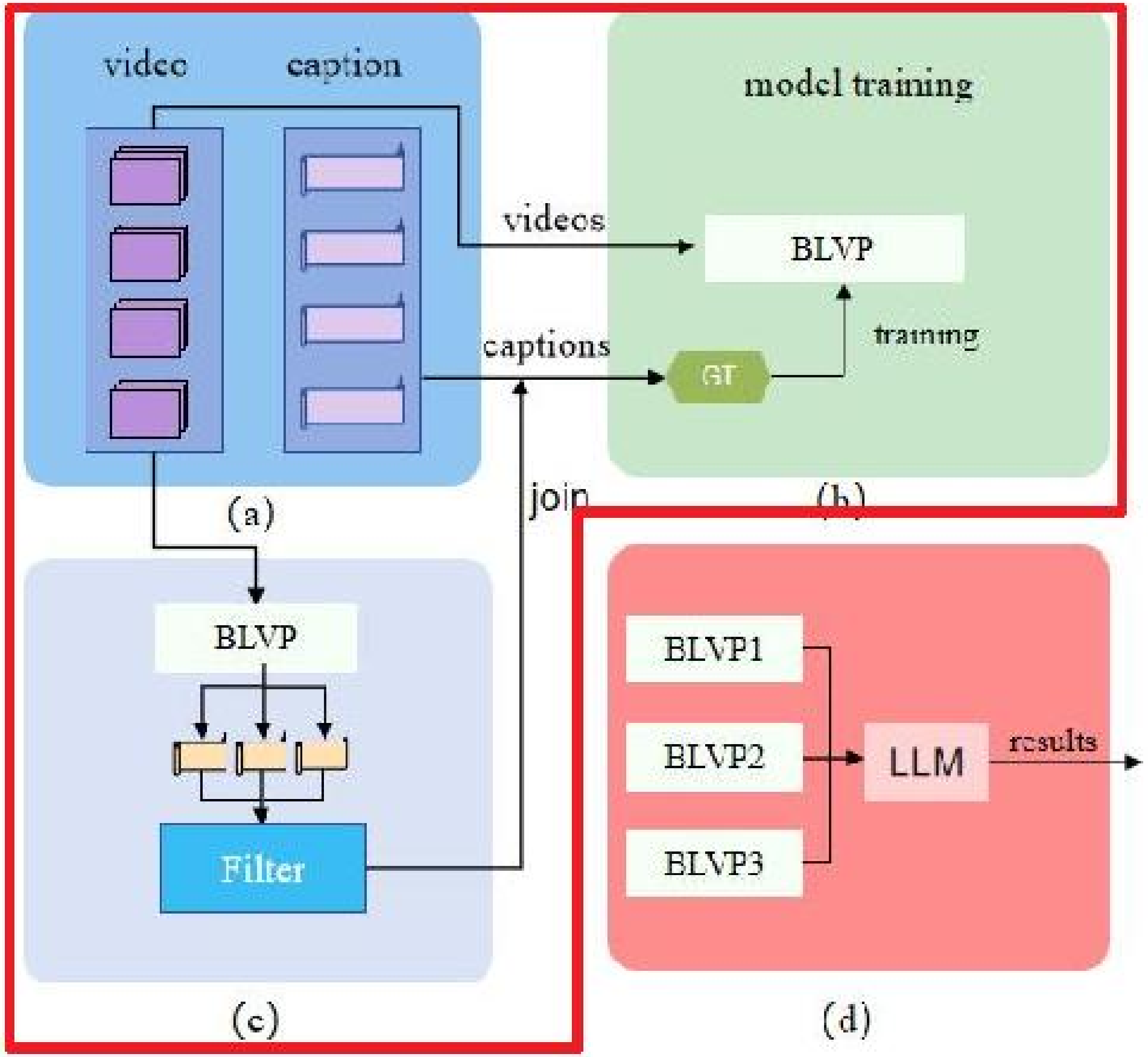
BLVP： BLIP2 for video to text

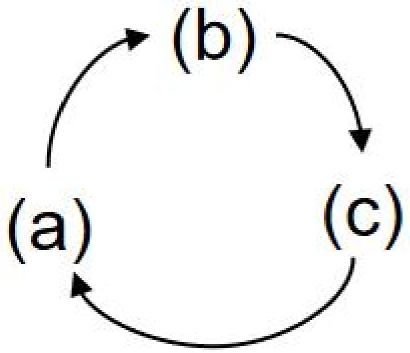


Video Encoder: CLIP-G

LLM Decoder: OPT 2.7

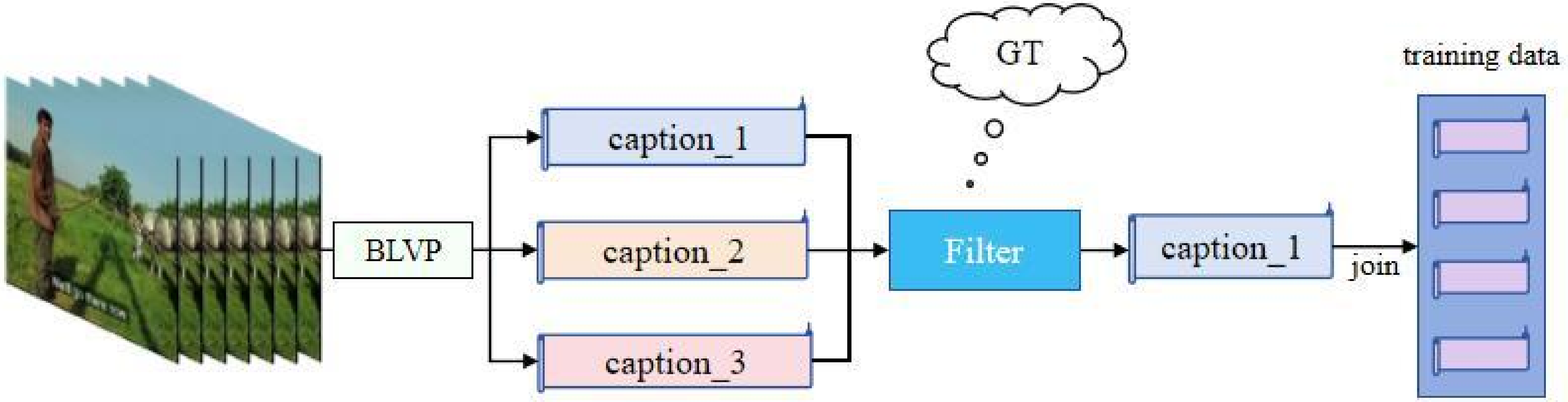
Cyclic data Augmentation





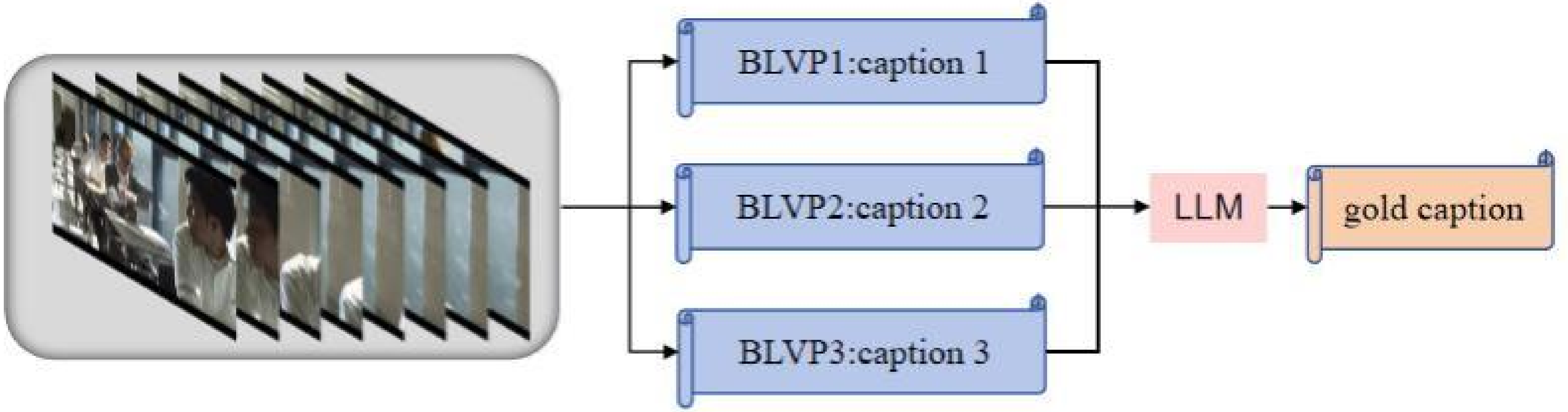
The cyclic workflow.

The process of data selection.



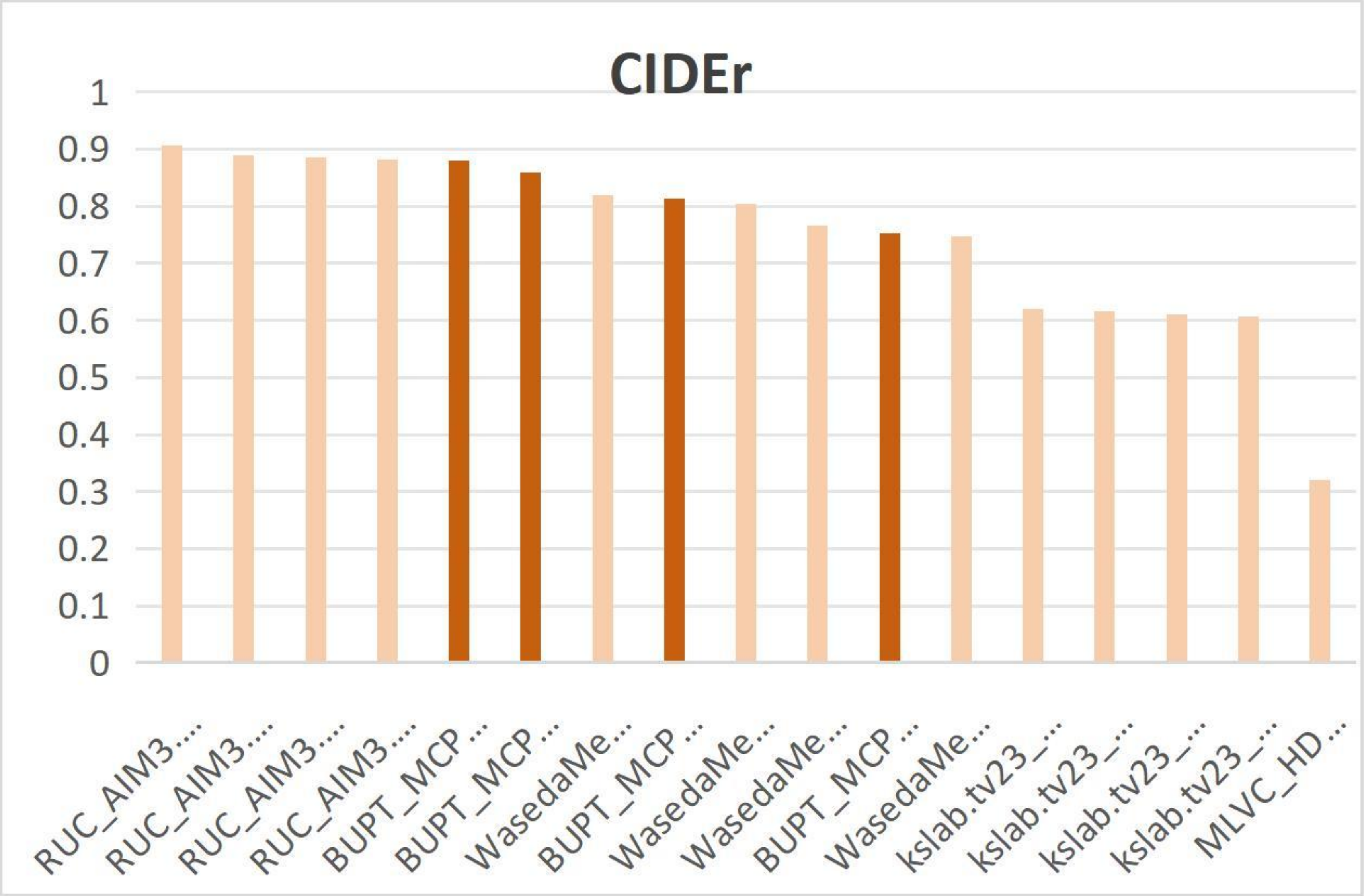
Filter rely on the CIDEr score calculated with respect to the Ground Truth.

LLM for Ensemble

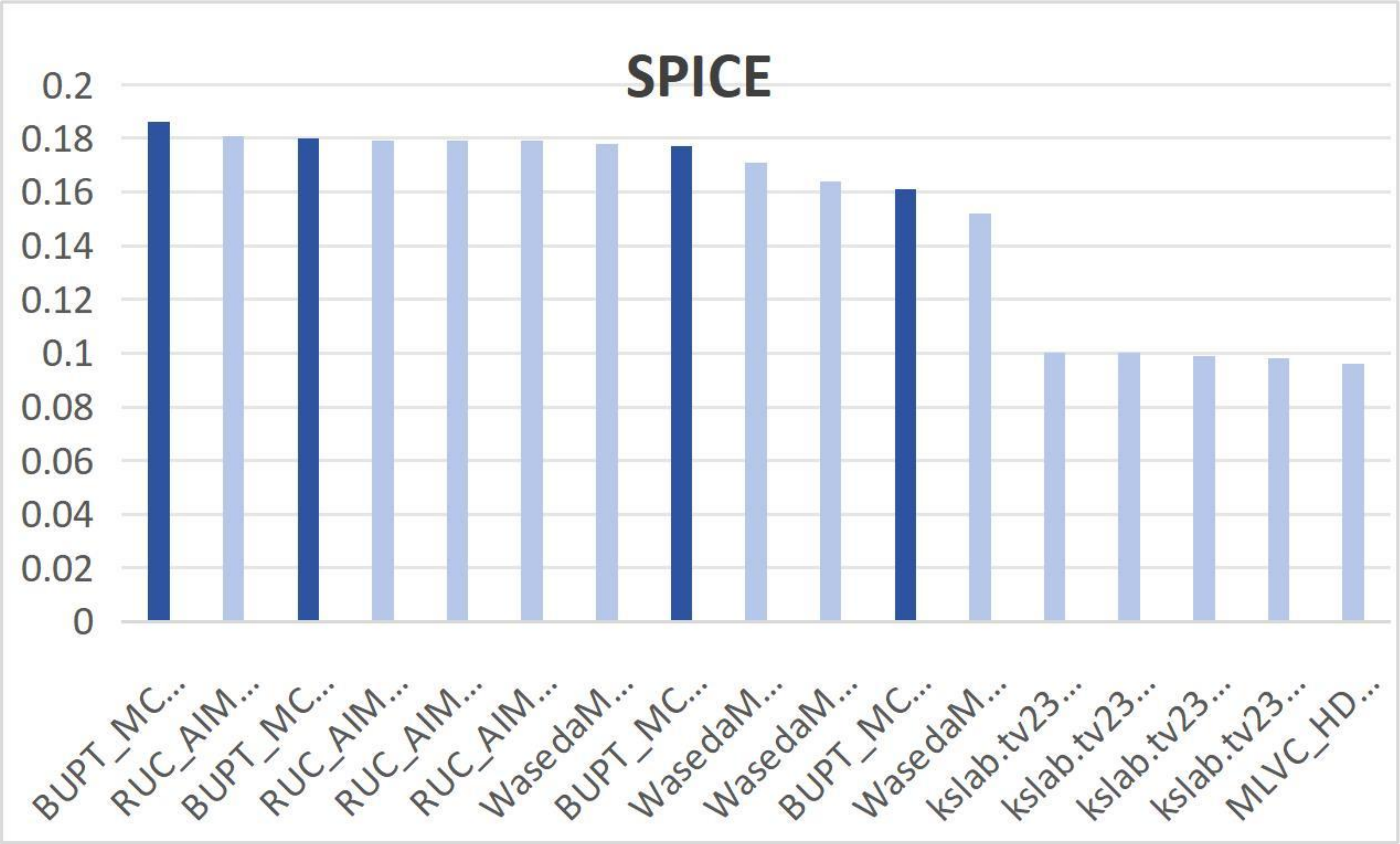


LLM: Alpaca-Lora 7B

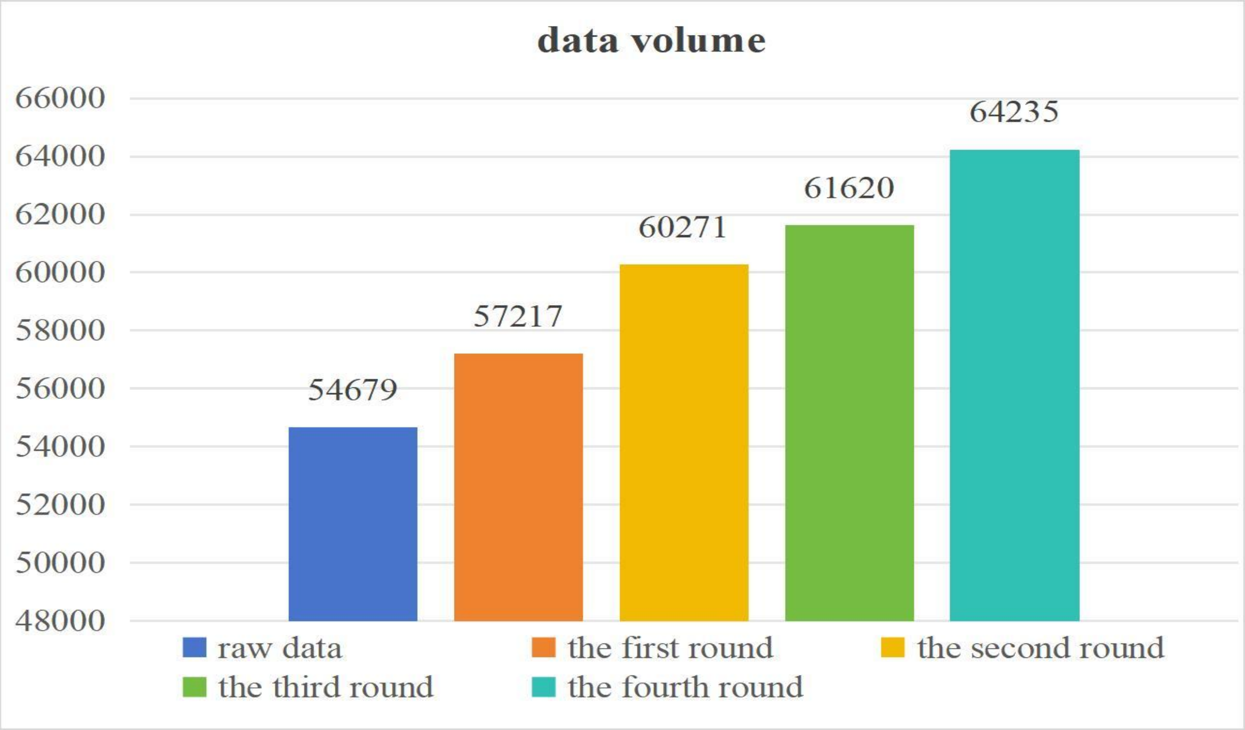
04 Results

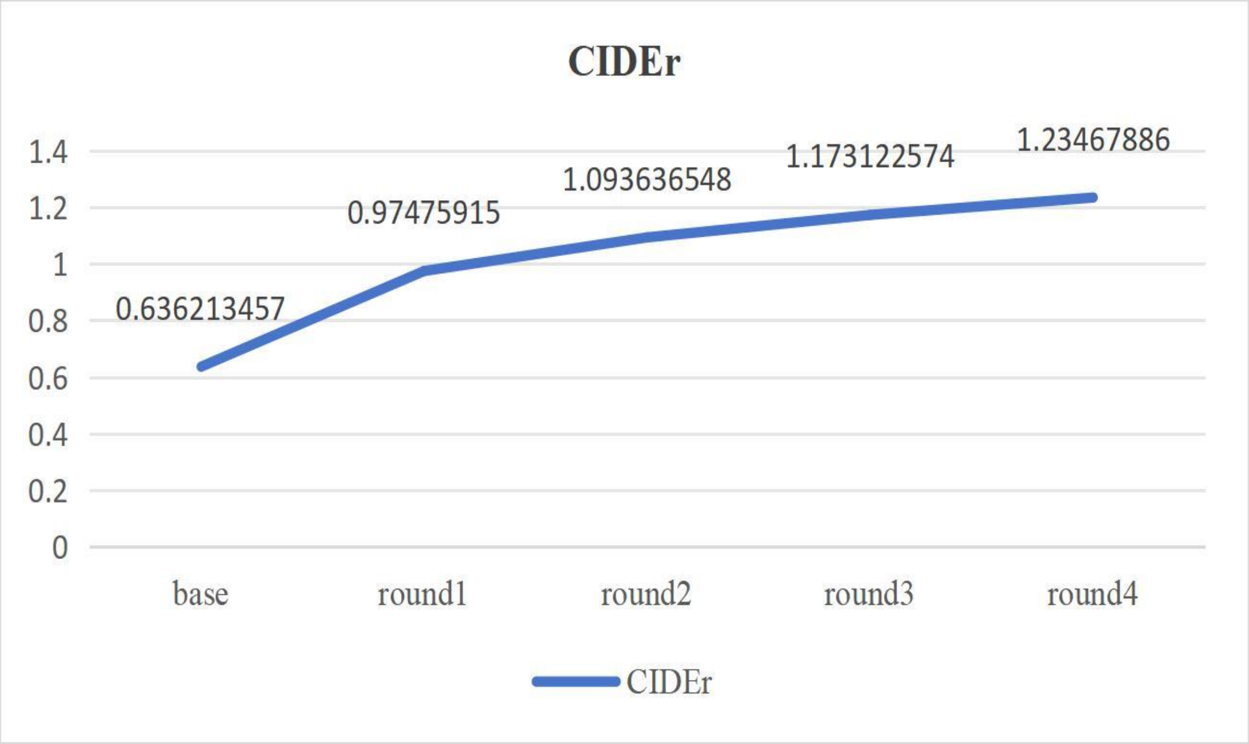

CIDEr
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
RUC_AIM3....
RUC_AIM3....
RUC_AIM3....
RUC_AIM3....
BUPT_MCP ...
BUPT_MCP ...
WasedaMe...
BUPT_MCP ...
WasedaMe...
WasedaMe...
BUPT_MCP ...
WasedaMe...
kslab.tv23_...
kslab.tv23_...
kslab.tv23_...
kslab.tv23_...
MLVC_HD...

04 Results


SPICE
0.2
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
BUPT_MC...
RUC_AIM...
BUPT_MC...
RUC_AIM...
RUC_AIM...
RUC_AIM...
WasedaM...
BUPT_MC...
WasedaM...
WasedaM...
BUPT_MC...
WasedaM...
kslab.tv23...
kslab.tv23...
kslab.tv23...
kslab.tv23...
MLVC_HD...

ablation experiments





04 Conclusion

In this study, we transferred the BLIP2 model to the VTT task and proposed a cyclic data augmentation approach. Our experimental results on the TRECVID VTT dataset achieved a CIDEr score of 87.9, ranking second in the competition. Additionally, to extract the semantic information, we used LLM to integrate the results of multiple models, obtaining a state-of-the-art SPICE evaluation metric.

However, there is room for improvement in our temporal modeling approach, particularly in understanding complex motion behaviors. Additionally, due to time constraints, we were only able to perform cyclic data augmentation, for a limited five rounds.

BUPT MCPRL

THANKS