Two-stage Ranking Strategy for Ad-hoc Video Search

Fangming Zhou[1], Changqiao Wu[2] , Xiaofeng Guo[2] , Haofan Wang[2] , Jincan Deng[2] , Debing Zhang[2]

[1]Renmin University of China

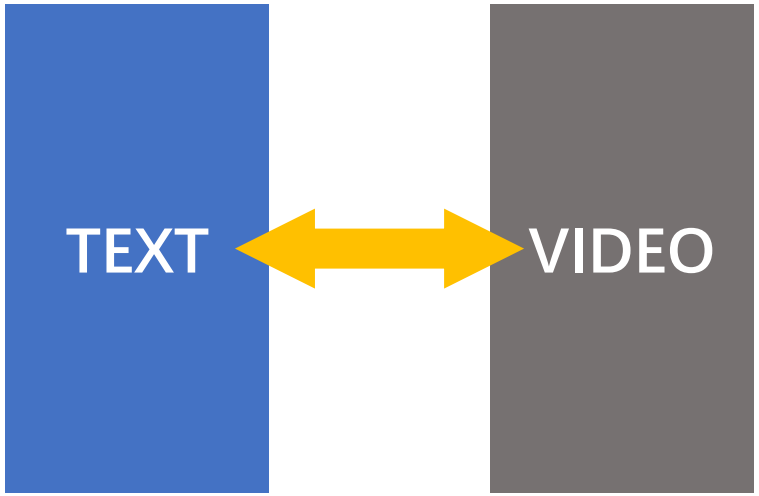
[2]MMU, Kuaishou Technology

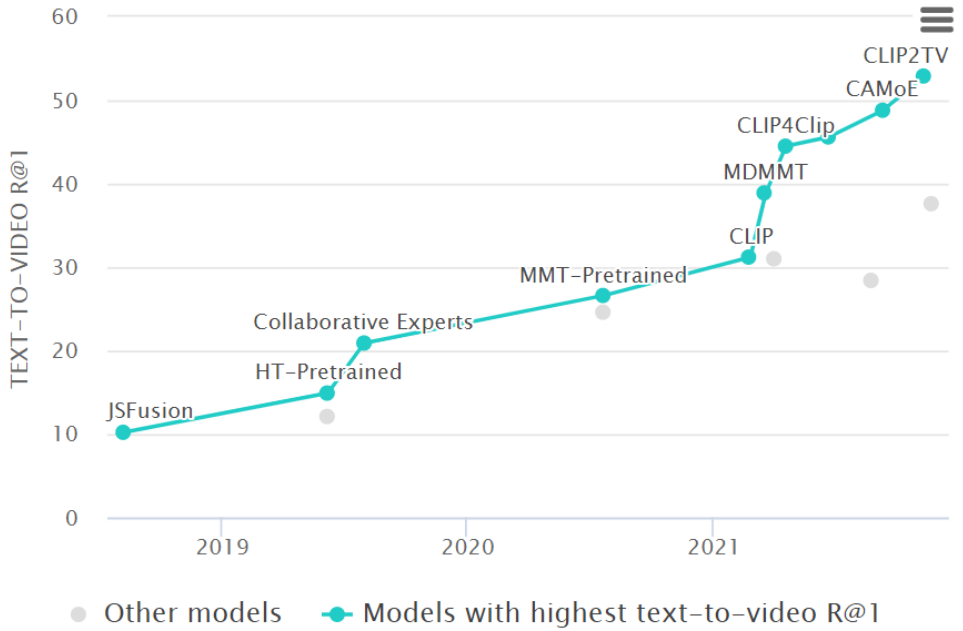
TRECVID 2021 workshop

2021-12

Ad-hoc Video Search

- **ESSENCE**: text-video matching in **retrieval** scenario (a hot topic)





Leaderboard of video retrieval on MSR-VTT-1kA [1]

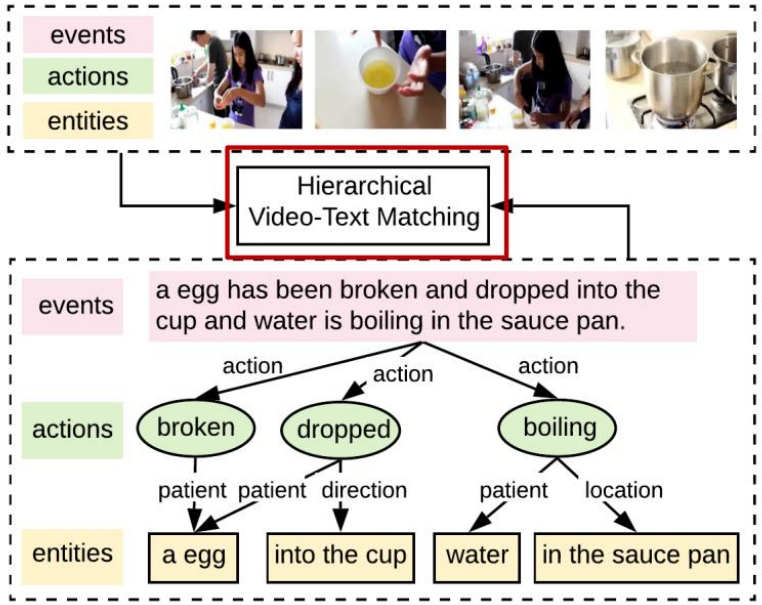
[1] MSR-VTT-1kA Benchmark (Video Retrieval) | Papers With Code

KUAISHOU

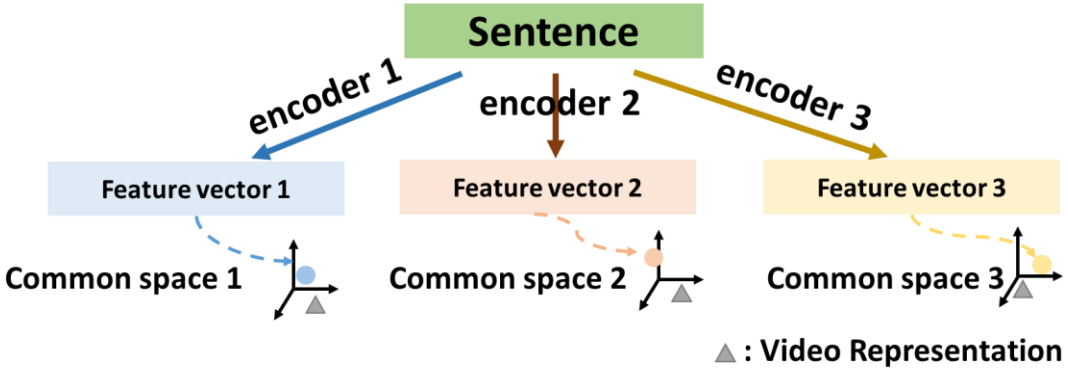
Previous works

- RUC_AIMC3 at TRECVID 2020 [1]:
  - Two-branch model
  - **Addition of irCSN feature**

- RUCMM at TRECVID 2020 [2]:
  - **Multi-space & multi-loss strategy**
  - Addition of C3d feature



RUC_AIMC3 at TRECVID 2020



RUCMM at TRECVID 2020

[1] Zhao et al., RUC_AIM3 at TRECVID 2020: Ad-hoc video search & video to text description. TRECVID, 2020.
[2] Li et al., Renmin University of China at TRECVID 2020: Sentence encoder assembly for ad-hoc video search. TRECVID, 2020.

KUAISHOU

Two-stage Ranking Strategy

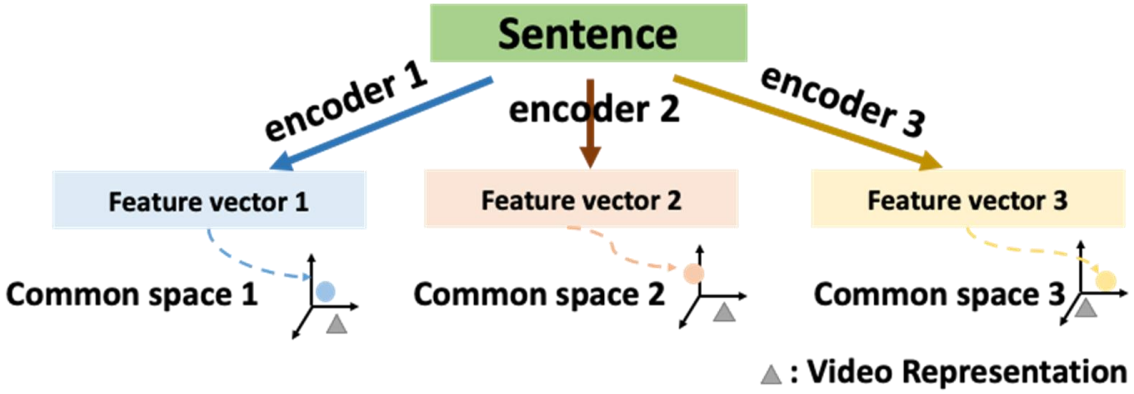
- **Stage - I : Keyword-based Rank**
  - Model architecture
  - Training data
  - Visual features and textual encoders

- **Stage - II : Fine-grained Re-rank**
  - Frame-level matching
  - Weighted sum of two stages as final similarity
  - Reasonableness of re-ranking

Two-stage Ranking Strategy

- Stage - I : Keyword-based Rank
  - **Model architecture**

SEA: sentence encoder assembly



SEA model [1]

- ✓ Multi-space architecture
- ✓ Learning k common space for k sentence encoders
- ✓ Combined loss:

$$loss = \sum_{i=1}^{k} loss_i(sentence, video)$$

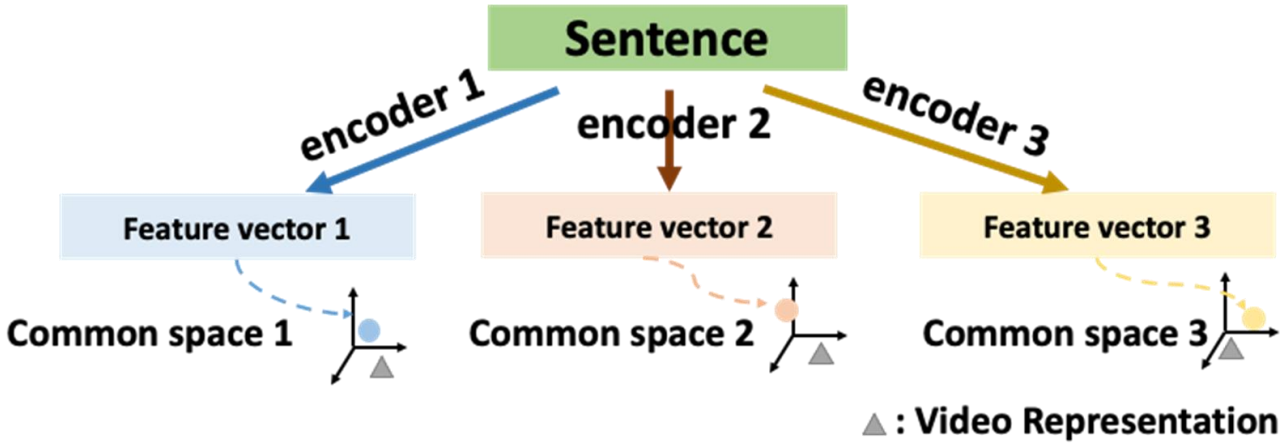
[1] Li et al., SEA: Sentence encoder assembly for video retrieval by textual queries. TMM, 2021.

KUAISHOU

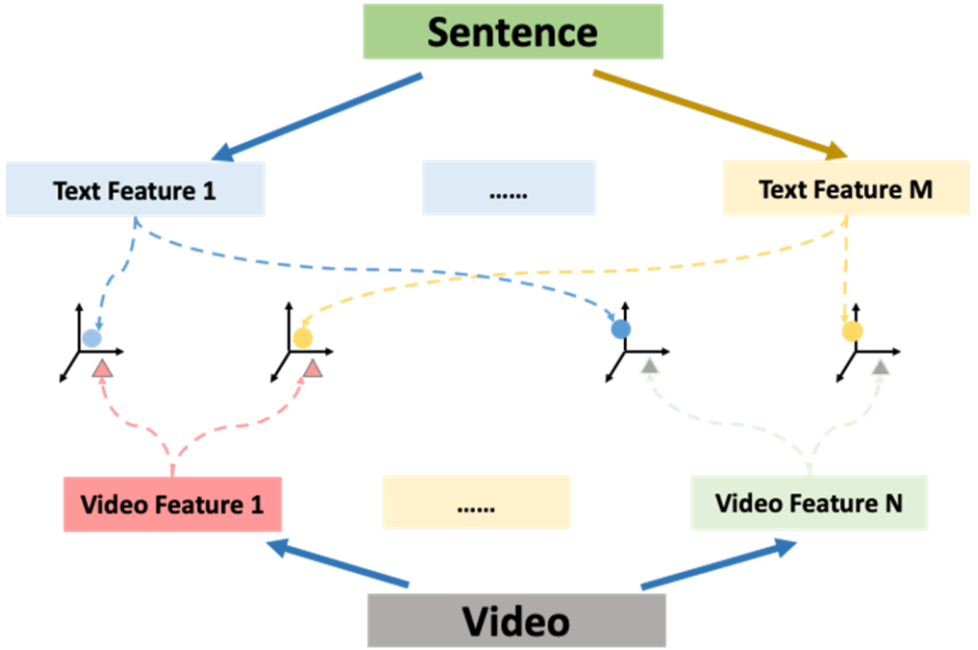
Two-stage Ranking Strategy

- Stage - I : Keyword-based Rank
  - **Model architecture**

SEA -> **SEA++**



SEA model [1]



SEA++ model

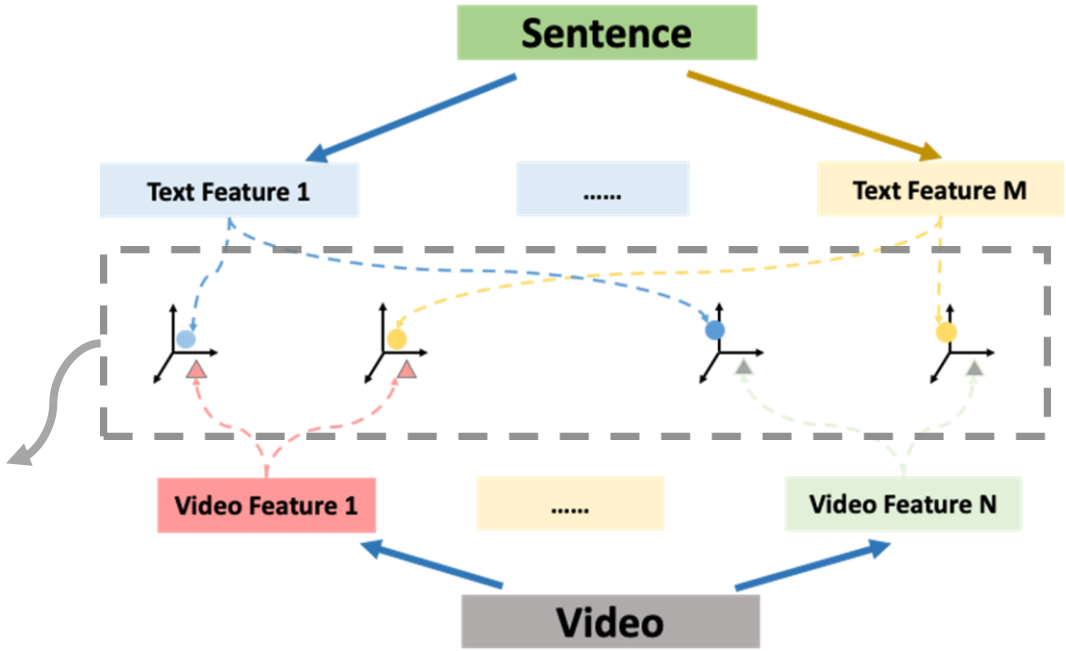
KUAISHOU

Two-stage Ranking Strategy

- Stage - Ⅰ : Keyword-based Rank
  - **Model architecture - SEA++ model**
    - ✓ Individual common space for each combination.
    - ✓ First stage similarity:

$$S_{first} = \sum_{i=1}^{m} \sum_{j=1}^{n} S_{space,i,j}$$

    - ✓ Only the first **K** videos sorted according to S_{first} will be passed to the Stage - Ⅱ .

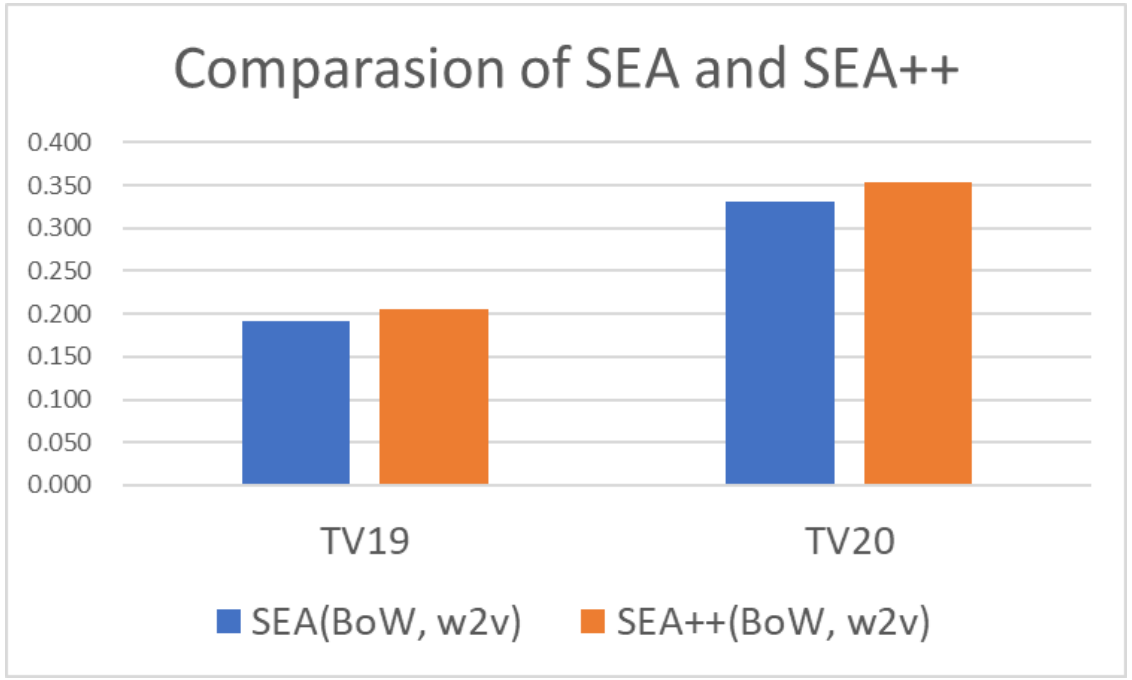


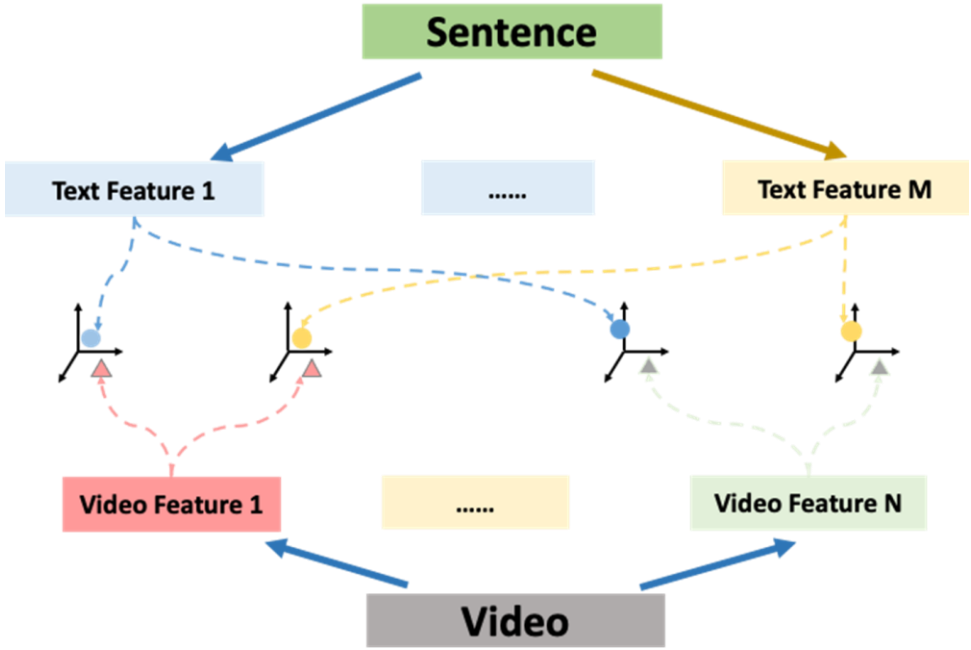
SEA++ model

KUAISHOU

Two-stage Ranking Strategy

- Stage - I : Keyword-based Rank
  - **Model architecture -** SEA++ model





SEA++ model

KUAISHOU

Two-stage Ranking Strategy

- Stage - I : Keyword-based Rank
  - **Training data**
    - ✓ Concepts in different datasets should be complementary.
    - ✓ Training data should be similar to V3C1.

Training data	Num of video/image	Num of sentence
MSR-VTT	10k	200k
TGIF	100k	124k
VATEX	32k	349k
MSCOCO	123k	616k

KUAISHOU

Two-stage Ranking Strategy

- Stage - I : Keyword-based Rank
  - **Visual features and textual encoders**

    ✓ **Visual features**: ResNeXt101, irCSN, CLIP, timesformer
    ✓ **Textual encoders**: Bag-of-word, word2vec (**keyword-based**)

Model	Feature	TV19	TV20
SEA(BoW, w2v)	Resnext+irCSN	0.167	0.316
	Resnext+irCSN+CLIP	0.185	0.327
	Resnext+irCSN+CLIP+timesformer	0.191	0.332

Ablation experiment on visual feature

KUAISHOU

Two-stage Ranking Strategy

- **Stage - II : Fine-grained Re-rank**
  - Frame-level matching
  - Weighted sum of two stages as final similarity
  - Reasonableness of re-ranking

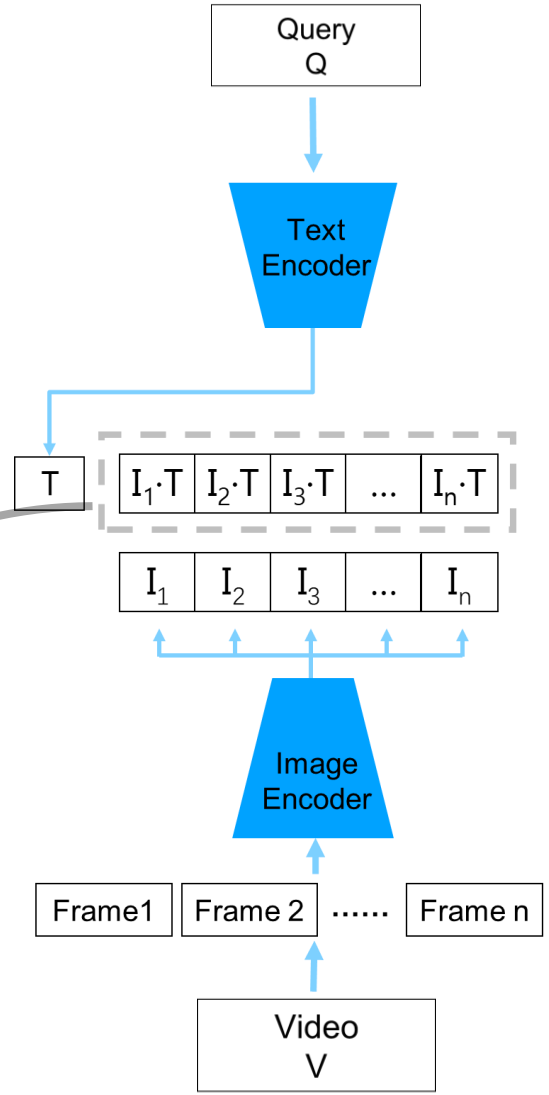
Two-stage Ranking Strategy

- Stage - II : Fine-grained Re-rank
  - **Frame-level matching**
    - ✓ We use out-of-box CLIP [1] as frame-level matching model.
    - ✓ Only Top-K videos sorted in Stage I are considered.
    - ✓ Second stage similarity:

$$S_{second} = \max(I_1 \cdot T, \cdots, I_n \cdot T)$$

$$I_i = ImageEncoder(frame_i)$$

$$T = TextEncoder(query)$$



[1] Radford et al,. Learning transferable visual models from natural language supervision, arxiv, 2021.

KUAISHOU

Two-stage Ranking Strategy

- Stage - II : Fine-grained Re-rank
  - **Weighted sum of two stages as final similarity**

$$S(Q,V) = \begin{cases} 0, & if\ S_{\boldsymbol{first}} < S_{threshold} \\ w_1 \cdot S_{\boldsymbol{first}} + w_2 \cdot S_{\boldsymbol{second}}, & if\ S_{\boldsymbol{first}} \geq S_{threshold} \end{cases}^{*}$$

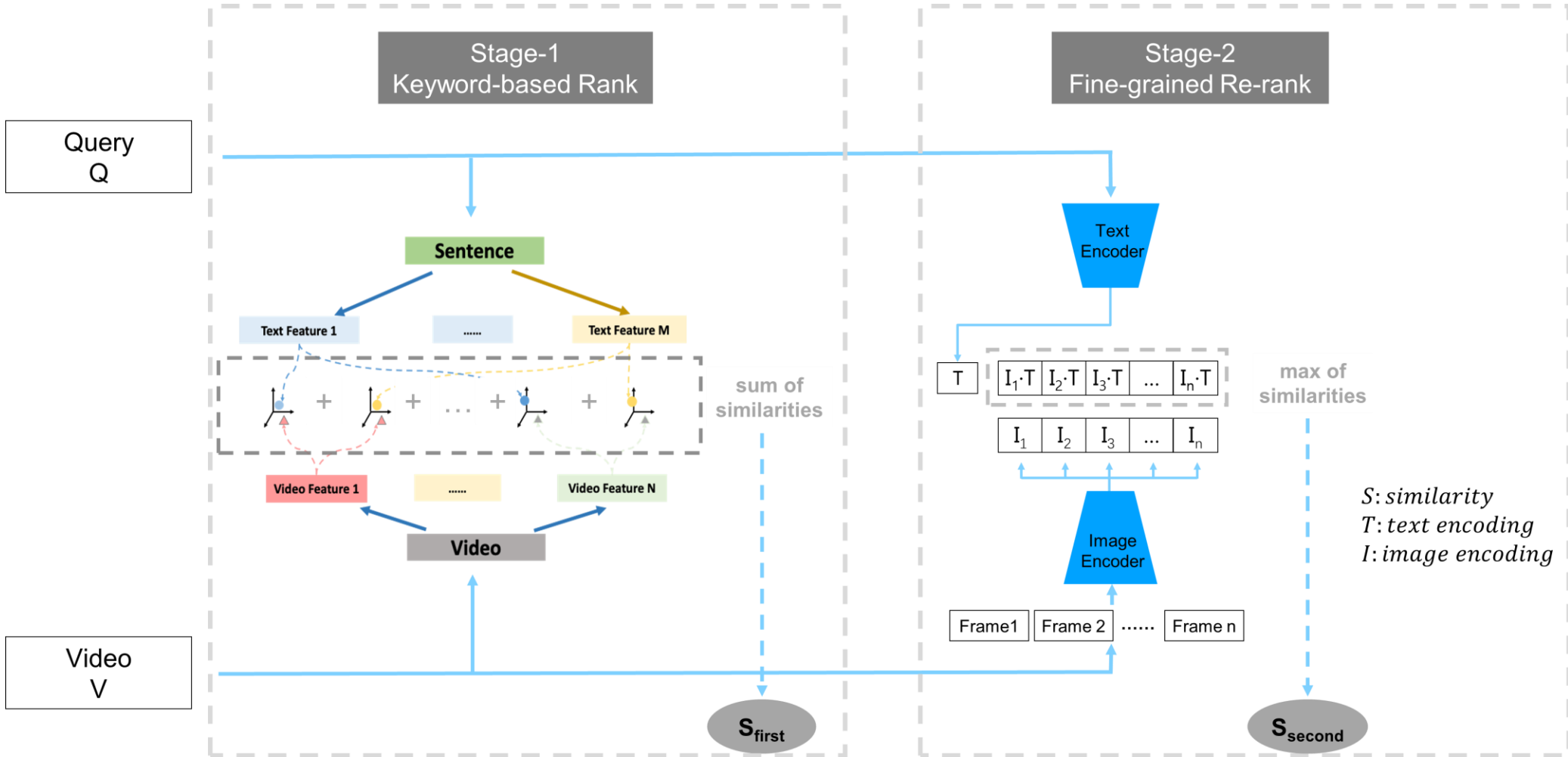
  - **Reasonableness of re-ranking**
    - ✓ Keyframe is enough in most cases.
    - ✓ The sentence semantics is considered (versus previous keyword-based method).

Models	TV19	TV20	TV21
baseline	0.211	**0.362**	0.340
Baseline + re-rank	**0.241**	0.360	**0.349**

[*] We set w_1=0.2 w_2=0.8 in our experiments.

KUAISHOU

Overall Pipeline



KUAISHOU

Submitted Runs

- Our final submitted runs as followed:
  - **run 3: single SEA++ model**

Submissions	TV19	TV20	TV21
Winner in 2019	0.163	-	-
Winner in 2020	-	0.359	-
run3	0.206	0.354	0.332

KUAISHOU

Submitted Runs

- Our final submitted runs as followed:
  - run 3: single SEA++ model
  - **run 2: model ensemble[1]**

Submissions	TV19	TV20	TV21
Winner in 2019	0.163	-	-
Winner in 2020	-	0.359	-
run3	0.206	0.354	0.332
run2	0.211	**0.362**	0.340

[1,2] indicates different ensemble strategy

KUAISHOU

Submitted Runs

- Our final submitted runs as followed:
  - run 3: single SEA++ model
  - run 2: model ensemble[1]
  - **run 1: model ensemble[1] + re-rank**

Submissions	TV19	TV20	TV21
Winner in 2019	0.163	-	-
Winner in 2020	-	0.359	-
run3	0.206	0.354	0.332
run2	0.211	**0.362**	0.340
run1(primary)	**0.241**	0.360	**0.349**

[1,2] indicates different ensemble strategy

KUAISHOU

Submitted Runs

- Our final submitted runs as followed:
  - run 3: single SEA++ model
  - run 2: model ensemble[1]
  - run 1: model ensemble[1] + re-rank
  - **run 4: model ensemble[2] + re-rank**

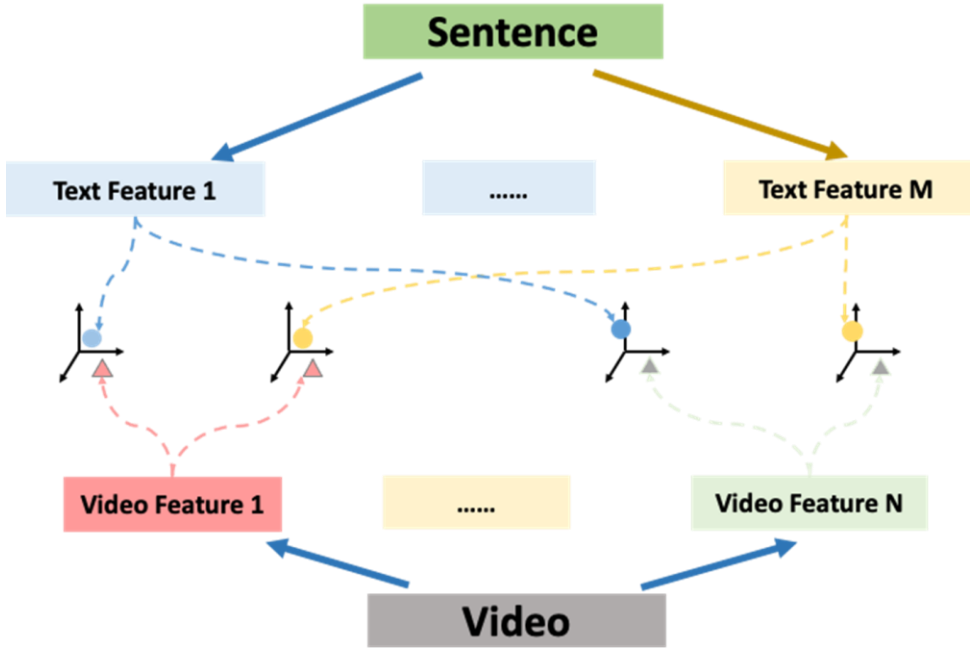
Submissions	TV19	TV20	TV21
Winner in 2019	0.163	-	-
Winner in 2020	-	0.359	-
run3	0.206	0.354	0.332
run2	0.211	**0.362**	0.340
run1(primary)	**0.241**	0.360	**0.349**
run4	0.239	0.358	**0.349**

[1,2] indicates different ensemble strategy

KUAISHOU

Take-home Message

- 1) We propose an improved video retrieval model, namely SEA++, which built a solid backbone for our best run.



SEA++ model

KUAISHOU

Take-home Message

- 1) We propose an improved video retrieval model, namely SEA++, which built a solid backbone for our best run.

- 2) Re-ranking by CLIP is an effective method to gain higher performance.

KUAISHOU

THANKS!

- **Contact with us**:

fangming_zhou@ruc.edu.cn,
wuchangqiao@kuaishou.com,
zhangdebing@kuaishou.com

KUAISHOU