Combining textual and visual features into multiple joint latent spaces, and introducing dual softmax re-ranking, for Ad-hoc Video Search

Damianos Galanopoulos, Vasileios Mezaris

Information Technologies Institute-CERTH

TRECVID 2022 Workshop

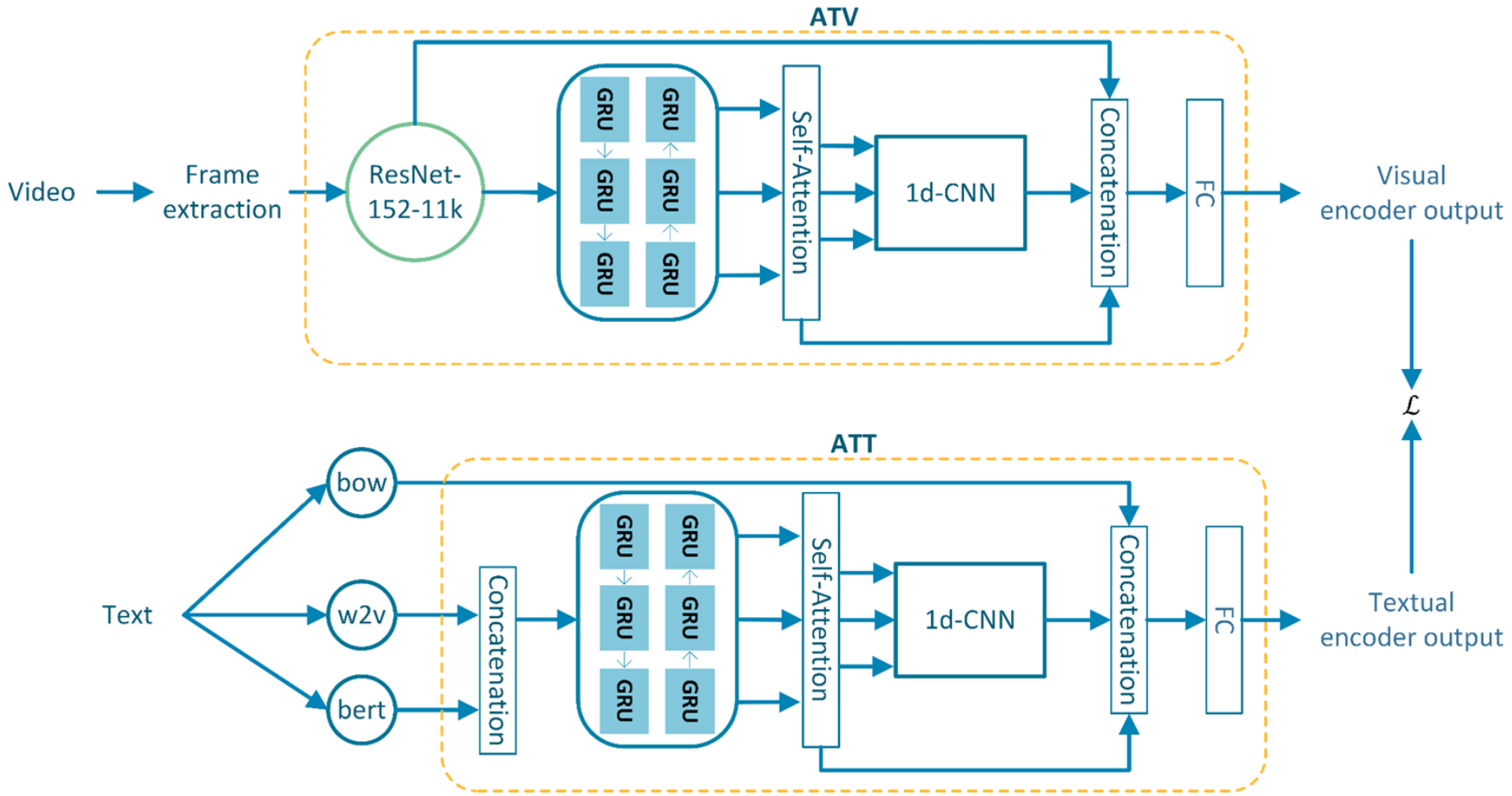
December 6 - 9, 2022

CRiTERIA - Comprehensive data-driven Risk and Threat Assessment Methods for the Early and Reliable Identification, Validation and Analysis of migration-related risks - has received funding from the European Union's Horizon 2020 research and innovation action program under grant agreement No 101021866.

Problem statement

- Retrieving video shots using natural-language textual queries
  - *“Find a shot of a man with a white beard”*

Our 2021 model

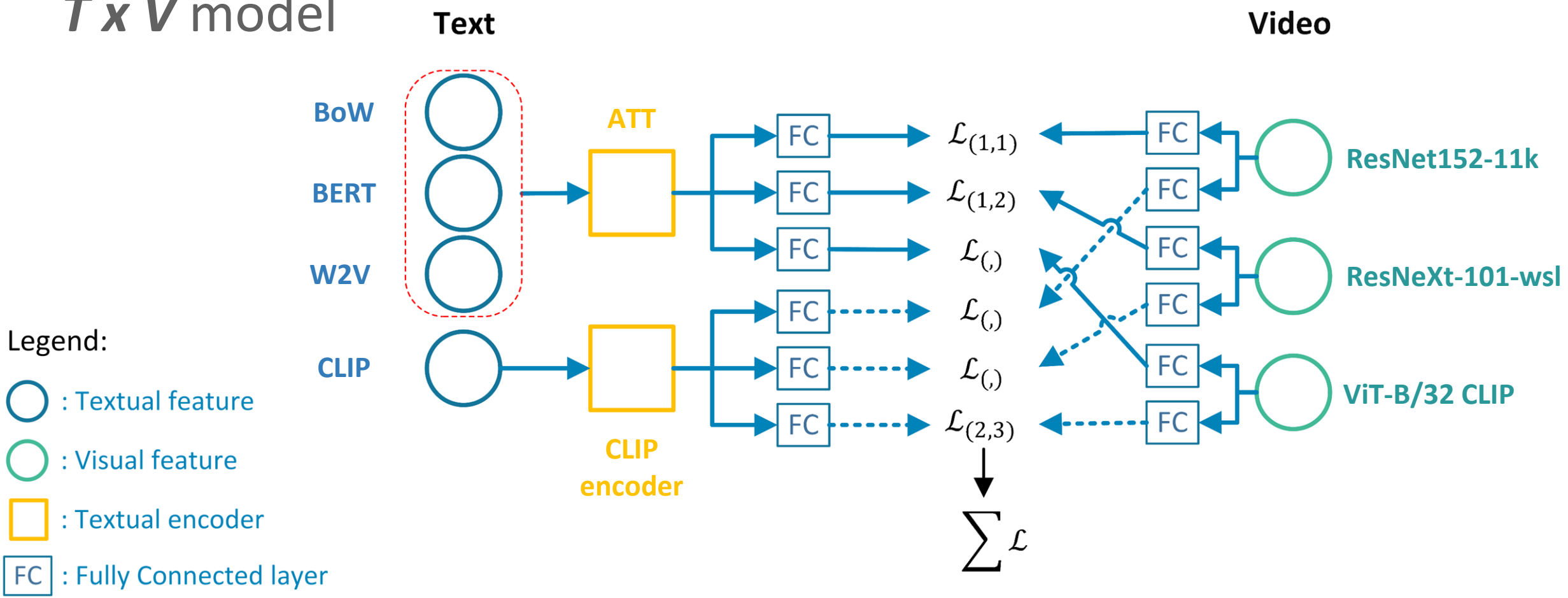


Our 2022 approach

- We utilize the T x V cross-modal network
- Combination of textual and visual features
- Multiple latent spaces development
- Multi-loss-based learning
- Similarity revision using a dual softmax operation

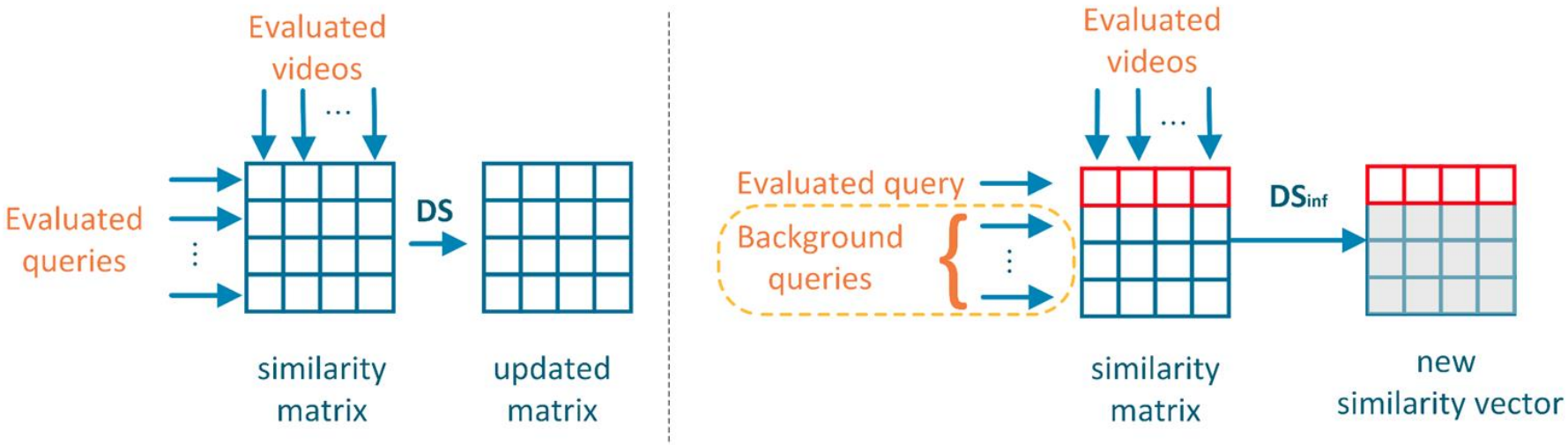
Our 2022 approach

T x V model



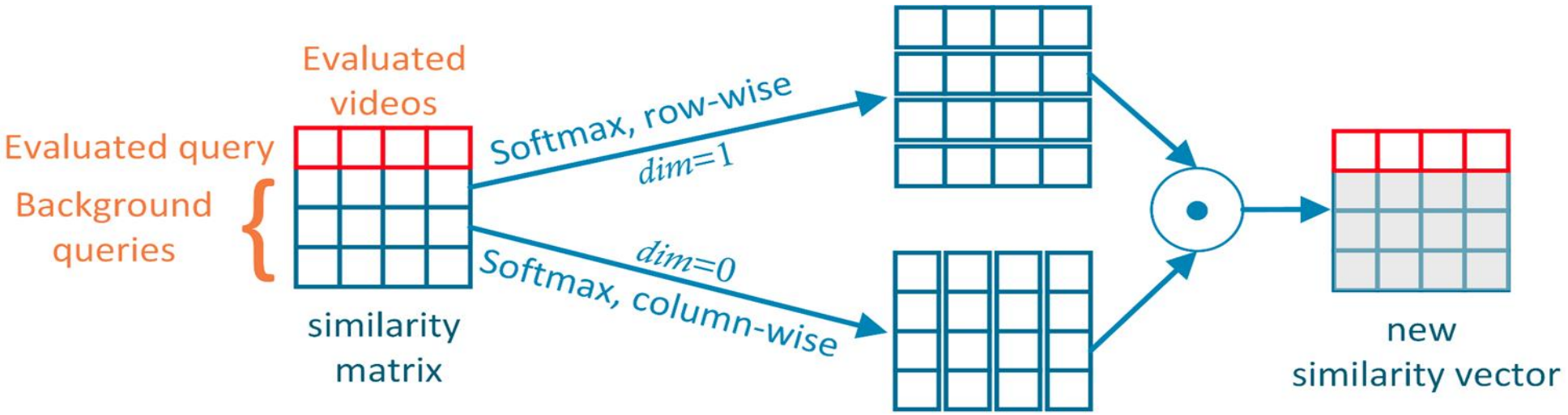
Dual Softmax Inference

- Query-video similarities revision
- Inspired by the Dual Softmax loss
- Two approaches:
  - Evaluation with *a priori* knowledge of all queries
  - Query-agnostic evaluation



Dual Softmax Inference

- Two different background queries strategies
  - AVS 2022 queries
  - AVS 2019-2020-2021 queries



Features and encoders

- Textual Features
  - Bag-of-words
  - BERT
  - Word2Vec
  - ViT-B/32 CLIP
- Textual Encoders
  - Attention-based dual encoding network (ATT)
  - ViT-B/32 CLIP encoder

Features and encoders

- Visual Features
  - **ResNet-152** trained on the ImageNet11k dataset
  - **ResNeXt-101** re-trained by weakly supervised learning on web images and fine-tuned on ImageNet
  - **ViT-B/32 CLIP** model

Datasets and metrics

- Datasets:
  - Training: MSR-VTT, TGIF, ActivityNet Captions and Vatex
  - Evaluation:
    - V3C2 evaluated on TRECVID AVS 2022 queries
    - V3C1 evaluated on TRECVID AVS 2019-2021 queries
    - IACC.3 evaluated on TRECVID AVS 2016-2018 queries
- Evaluation metric:
  - Mean extended inferred average precision (MXinfAP)

Submitted runs

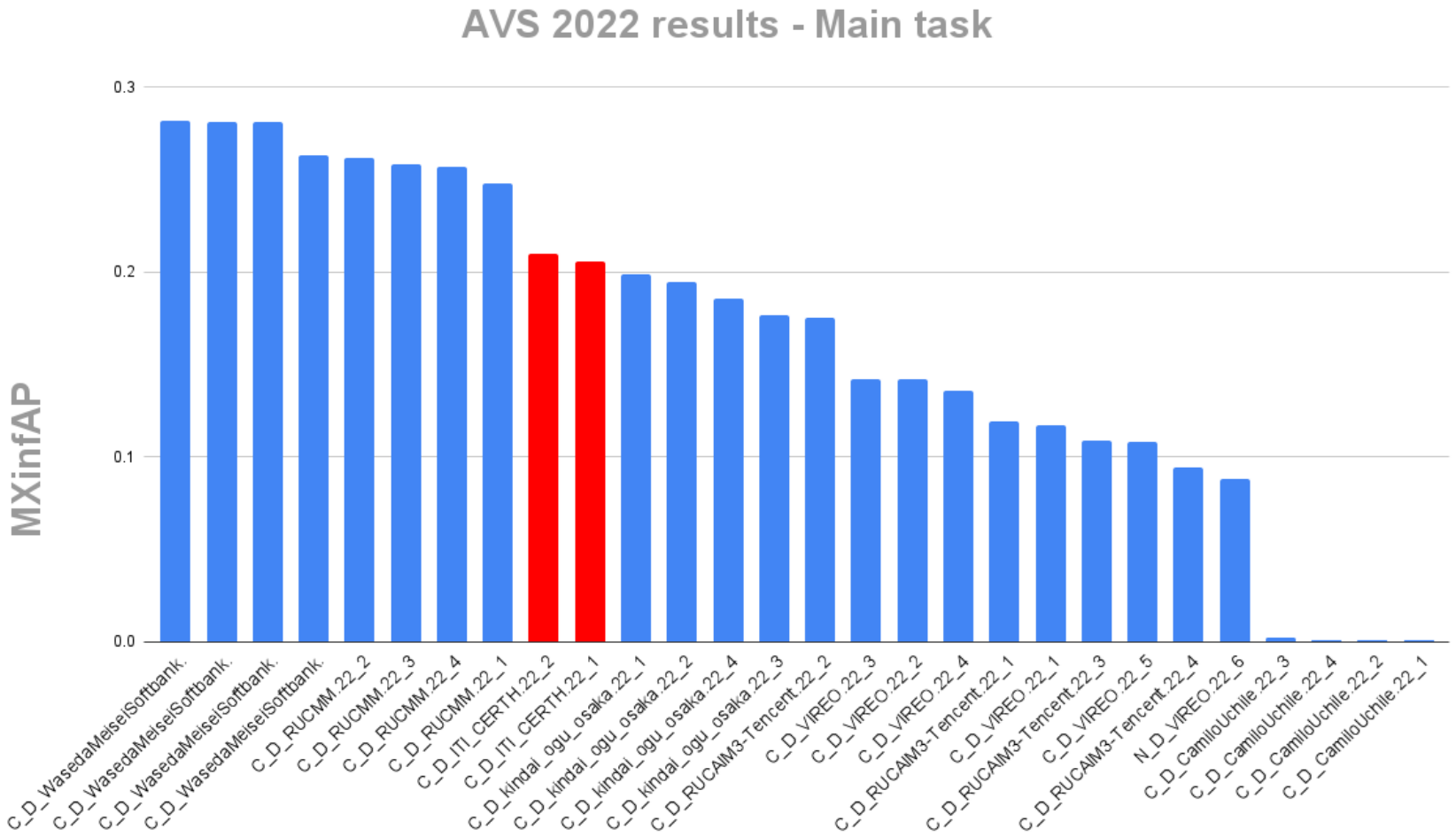
- **ITI_CERTH.22_run_2**:
  - ***T x V*** model with 2 textual encoders and 3 visual features
  - Late fusion of 6 trained models with different configurations
  - Dual softmax using all AVS2022 queries as background queries
- **ITI_CERTH.22_run_1:**
  - Similar to run #2
  - Dual softmax using AVS 2019-2021 queries as background

Results

Model	Dataset						
	IACC.3			V3C1			V3C2
	AVS16	AVS17	AVS18	AVS19	AVS20	AVS21	AVS22
T×V	0.234	0.317	0.153	0.220	0.316	0.312	0.194
T×V + DS on the set of evaluated queries (run #2)	**0.244**	**0.330**	**0.165**	**0.226**	0.324	**0.324**	**0.210**
T×V + DS using other years' queries of the same dataset (run #1)	0.243	0.328	0.162	**0.226**	**0.325**	0.323	0.206

MXinfAP for all submitted runs and TxV baseline for the fully-automatic AVS task.

Results



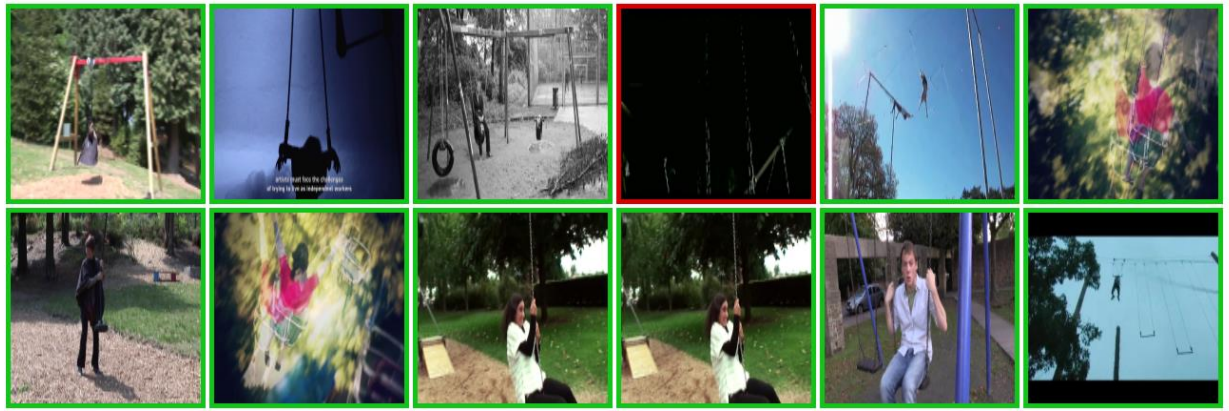
AVS 2022 ranking list of all submitted runs regarding the main task in MXinfAP terms. Red bars indicate our submitted runs.

CRiTERIA

Dual Softmax impact

“A person is in the act of swinging”

T x V model



T x V model with DS (run #2)

Dual Softmax impact

“A drone landing or rising from the ground”

T x V model

T x V model with DS (run #2)

Dual Softmax impact

“A type of cloth hanging on a rack, hanger, or line”

T x V model

T x V model with DS (run #2)

Conclusions

- We utilize the cross-modal TxV network
- We combine multiple textual and visual features
- The dual softmax operation boosts the performance
- The *a priori* knowledge of evaluation queries benefits the performance
- A query-agnostic dual softmax revision serves real-life scenarios, with a small trade-off regarding the overall performance

Thank you!

Damianos Galanopoulos, Vasileios Mezaris

Information Technologies Institute-CERTH

dgalanop@iti.gr, bmezaris@iti.gr

www.iti.gr/~bmezaris