UCF Center for Research in Computer Vision

NEURAL INFORMATION PROCESSING SYSTEMS

From Play to Replay: Composed Video Retrieval for Temporally Fine-Grained Videos

Animesh Gupta
University of Central Florida, USA

Jay Parmar
University of Central Florida, USA

Ishan Dave
Adobe Research, USA

Mubarak Shah
University of Central Florida, USA

UCF

What is Composed Video Retrieval?

Given a query video and modification, retrieve the intended video.

Query Video

"River landscape in **autumn**"

Modification

change the season to **springtime**

Retrieve

"River landscape in **springtime**"

Existing Work

WebVid [1]

Limitation: Appearance changes and minimal temporal reasoning.

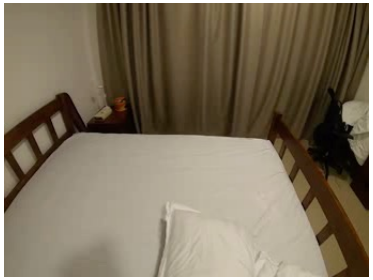
River landscape in **springtime**

change the season to **springtime**

River landscape in **autumn**

EgoCVR [2]

Limitation: Query and target from different segments of the same video.



#C C **puts** pillow **down**

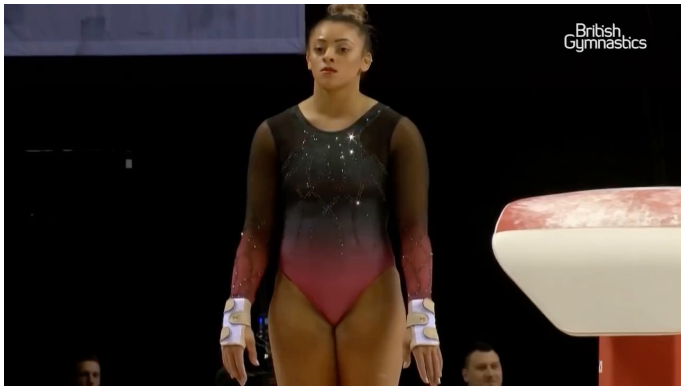
Pick it **up**

#C C **picks** pillow

[1] Lucas Ventura, Antoine Yang, Cordelia Schmid, and Gul Varol. CoVR: Learning Composed Video Retrieval from Web Video Captions. In AAAI 2024
[2] Thomas Hummel, Shyamgopal Karthik, Mariana-Iuliana Georgescu, and Zeynep Akata. EgoCVR: An Egocentric Benchmark for Fine-Grained Composed Video Retrieval. In ECCV 2024.

Temporally Fine Grained – Composed Video Retrieval (TF-CoVR)

Query Video



(Vault) round-off, flic-flac with 0.5 turn on, stretched salto forward with **0.5 turn off**

Modification

show with **2 turn**

Target Video

(Vault) round-off, flic-flac with 0.5 turn on, stretched salto forward with **2 turn off**

TF-CoVR: Multiple Ground-Truths

Query Video

Description

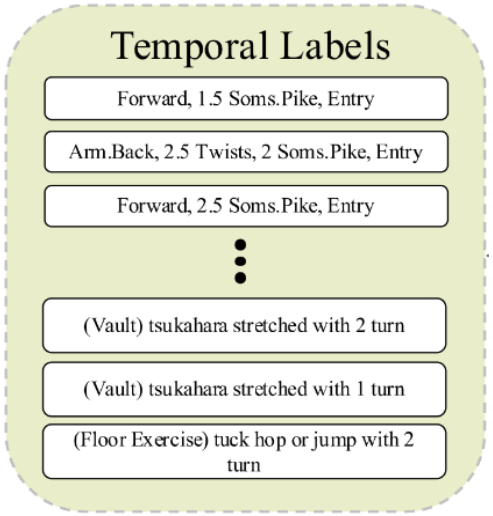
Inward, **3.5 Soms.**Tuck

Modification

Show with **2.5 somersaults**

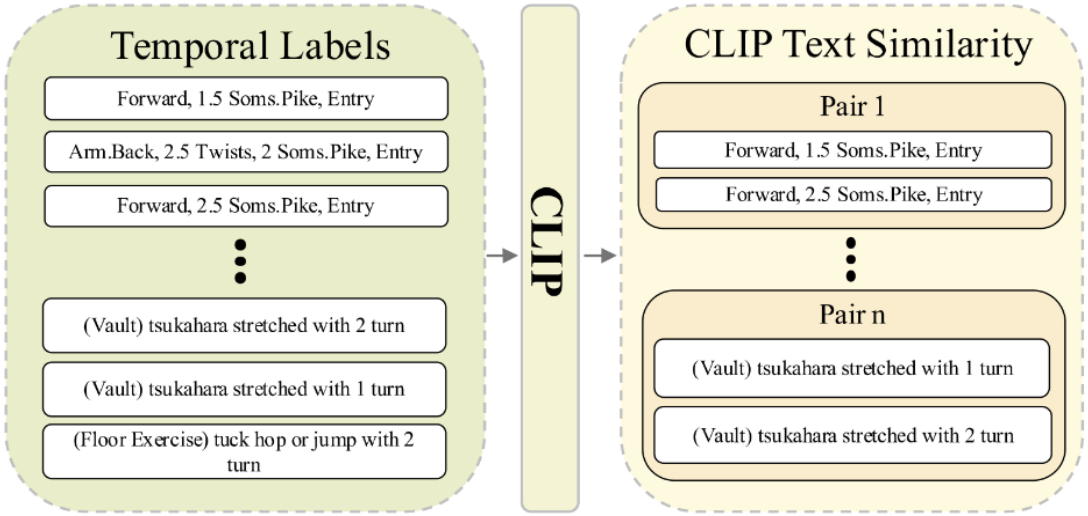
Rank 1 | Rank 2 | Rank 3 | Rank 4 | Rank 5

TF-CoVR: Dataset Generation



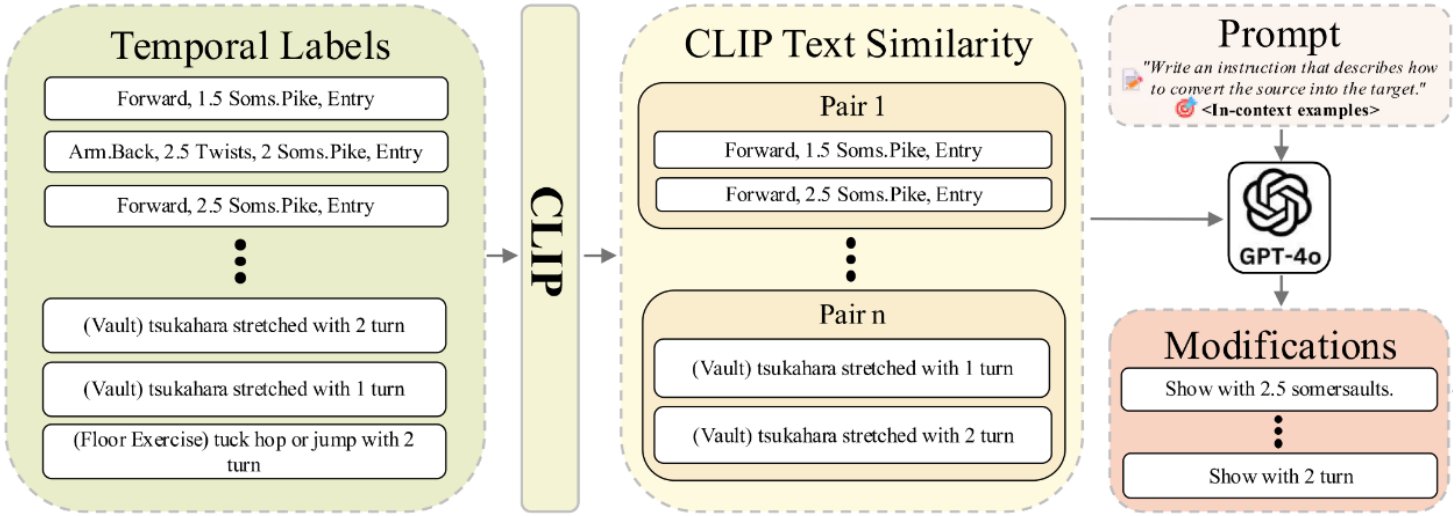
- Extract temporal action labels from *FineGym* (288 labels) and *FineDiving* (52 labels) datasets.

TF-CoVR: Dataset Generation



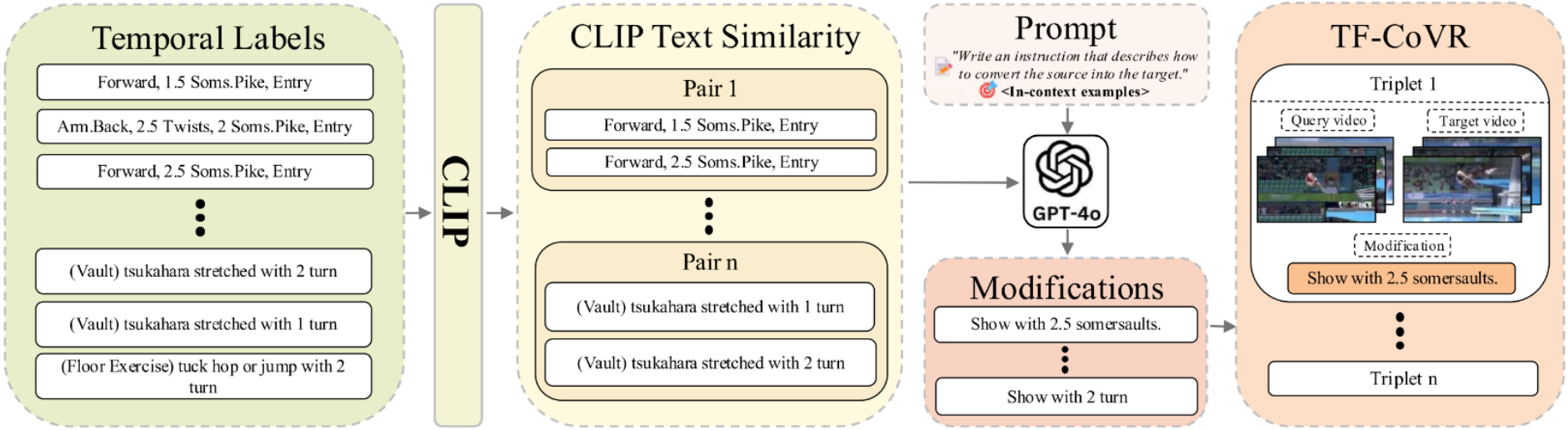
- Extract temporal action labels from *FineGym* (288 labels) and *FineDiving* (52 labels) datasets.
- Compute CLIP similarity between labels to generate *query-target* text pairs.

TF-CoVR: Dataset Generation



- Extract temporal action labels from *FineGym* (288 labels) and *FineDiving* (52 labels) datasets.
- Compute CLIP similarity between labels to generate *query-target* text pairs.
- Use GPT-4o with prompts and in-context examples to produce *modification texts*.

TF-CoVR: Dataset Generation

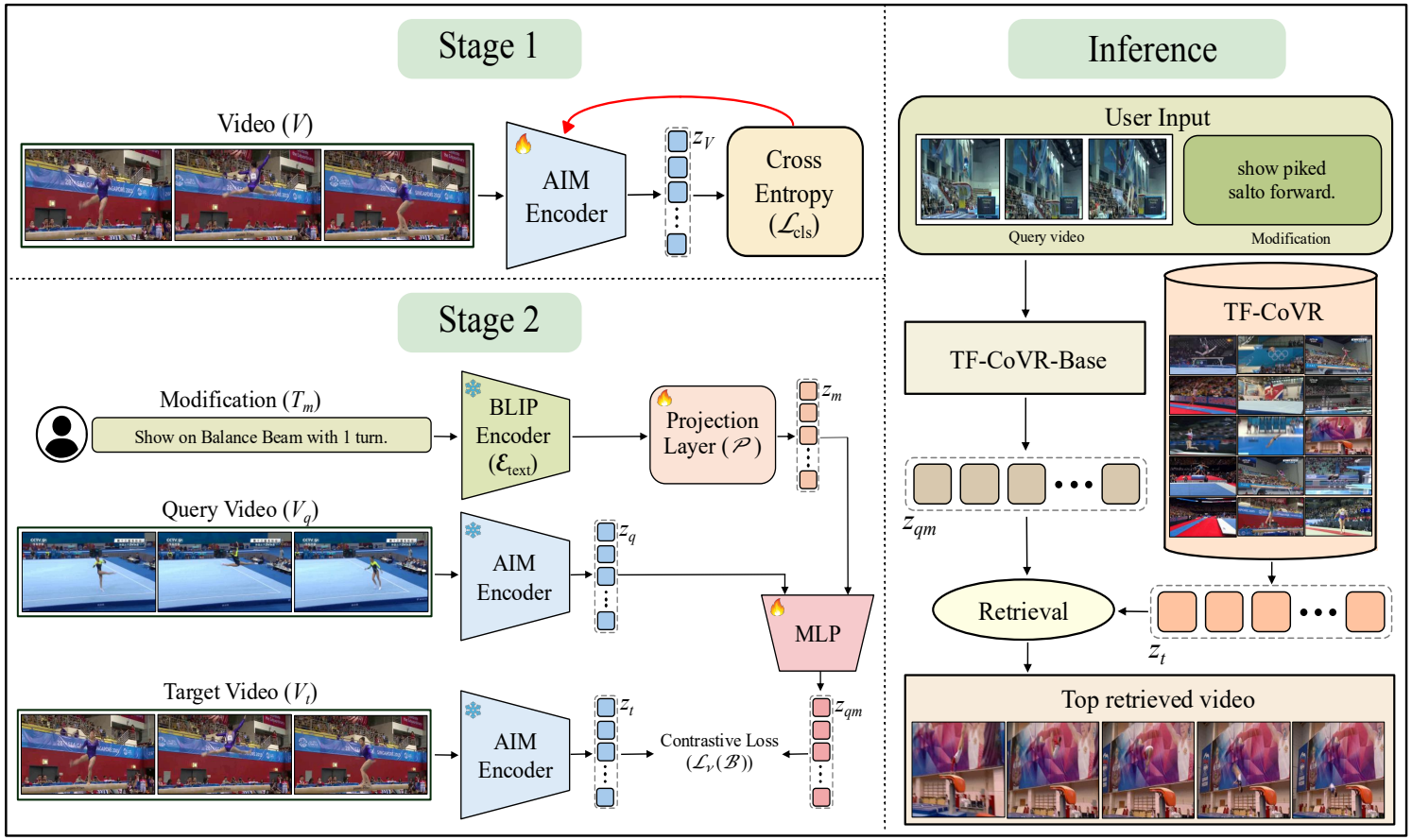


- Extract temporal action labels from *FineGym* (288 labels) and *FineDiving* (52 labels) datasets.
- Compute CLIP similarity between labels to generate *query-target* text pairs.
- Use GPT-4o with prompts and in-context examples to produce *modification texts*.
- *TF-CoVR* triplets consist of <*query video, target video, modification text*>.

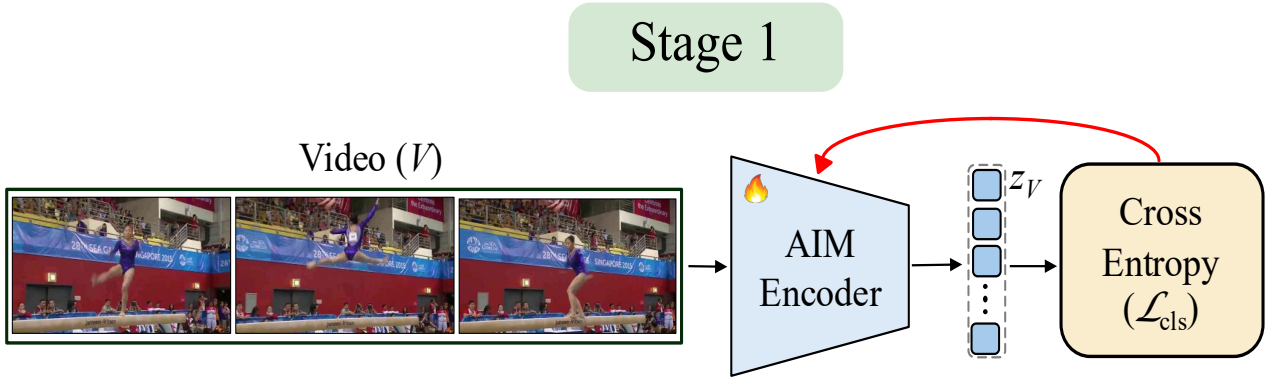
TF-CoVR: Comparison

Dataset	Type	#Triplets	Train	Eval	Multi-GT	Eval Metrics	#Sub-actions
CIRR [24]	📷	36K	✓	✓	✗	Recall@K	✗
FashionIQ [41]	📷	30K	✓	✓	✗	Recall@K	✗
CC-CoIR [38]	📷	3.3M	✓	✗	✗	Recall@K	✗
MTCIR [12]	📷	3.4M	✓	✗	✗	Recall@K	✗
WebVid-CoVR [38]	🎥	1.6M	✓	✓	✗	Recall@K	✗
EgoCVR [9]	🎥	2K	✗	✓	✗	Recall@K	✗
FineCVR [47]	🎥	1M	✓	✓	✗	Recall@K	✗
CIRCO [3]	📷	800	✗	✓	✓	mAP@K	✗
TF-CoVR (Ours)	🎥	180K	✓	✓	✓	mAP@K	306

TF-CoVR-Base: Method

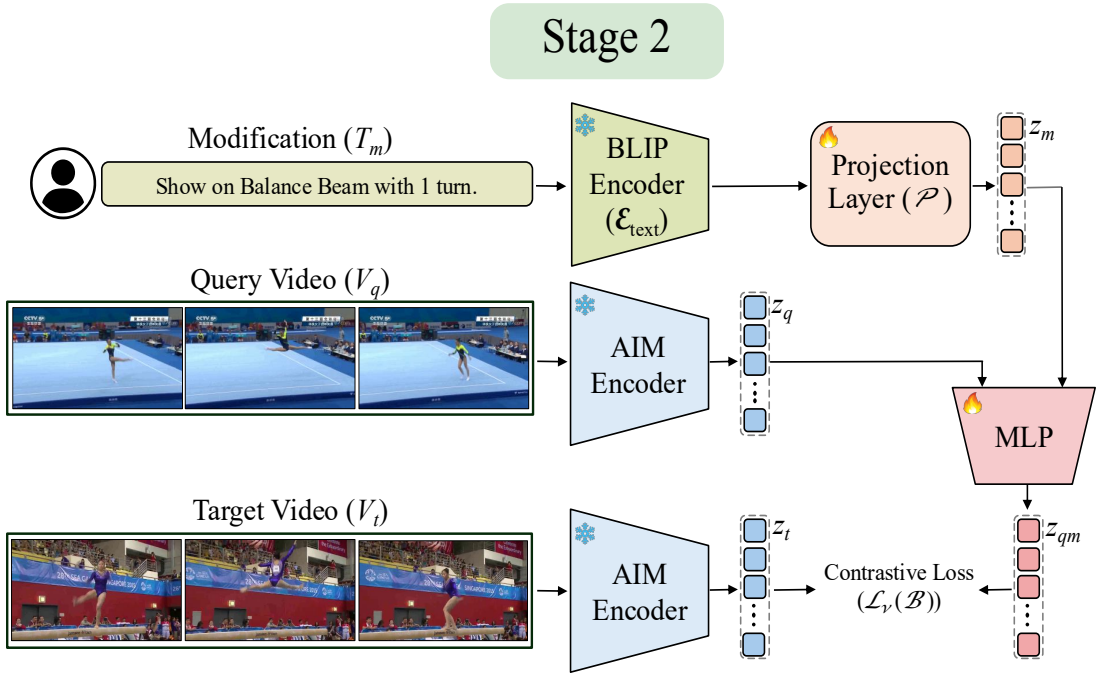


TF-CoVR-Base: Stage 1



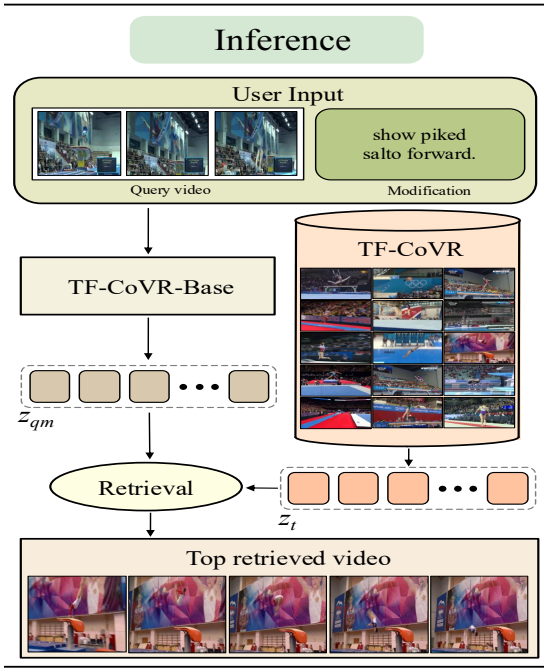
- Pretrain AIM encoder on all videos to capture fine-grained motion.
- Train on 306 fine-grained action classes with cross-entropy loss to learn temporal discrimination.

TF-CoVR-Base: Stage 2



- BLIP and AIM encoders extract embeddings from the *modification text* (z_m) and *query video* (z_q).
- An MLP fuses z_q and z_m into a composed query embedding (z_{qm}).
- z_{qm} is aligned with *target video* embedding (z_t) using hard negative-aware contrastive loss (HN-NCE).

TF-CoVR-Base: Inference



- At inference, the query video and text are fused using finetuned AIM, BLIP and MLP to create a composed query embedding.
- This is then compared with all TF-CoVR video embeddings using cosine similarity to rank and retrieve the best match.

Results: Pre-trained

Modalities		Model	Fusion	#Query	#Target	mAP@K (↑)			
Video	Text			Frames	Frames	5	10	25	50
General Multimodal Embeddings (TF-CoVR)									
✓	✓	GME-Qwen2-VL-2B [51]	MLLM	1	15	2.28	2.64	3.29	3.81
✓	✓	MM-Embed [19]	MLLM	1	15	2.39	2.81	3.61	4.14
✓	✓	E5-V [15]	Avg	1	15	3.14	3.78	4.65	5.22
Not fine-tuned on TF-CoVR									
✗	✓	BLIP2	-	-	15	1.34	1.79	2.20	2.50
✓	✗	BLIP2	-	1	15	1.74	2.20	3.06	3.62
✓	✓	BLIP-CoVR [41]	CA	1	15	2.33	2.99	3.90	4.50
✓	✓	$\text{BLIP}_{\text{CoVR-ECDE}}$ [37]	CA	1	15	0.78	0.88	1.16	1.37
✗	✓	TF-CVR [9]	-	-	15	0.56	0.76	0.99	1.24
✓	✓	LanguageBind [55]	Avg	8	8	3.43	4.37	5.26	5.92

Results: Fine-tuned

Modalities		Model	Fusion	#Query	#Target	mAP@K (↑)			
Video	Text			Frames	Frames	5	10	25	50
Fine-tuned on TF-CoVR									
✗	✓	BLIP2	-	-	15	10.69	13.02	15.35	16.41
✓	✗	BLIP2	-	1	15	4.86	6.49	8.92	10.06
✓	✓	CLIP	MLP	1	15	7.01	8.35	10.22	11.38
✓	✓	BLIP2	MLP	1	15	10.86	13.20	15.38	16.31
✓	✓	CLIP	MLP	15	15	6.40	7.46	9.21	10.40
✓	✓	BLIP2	MLP	15	15	11.64	14.81	16.74	17.55
✓	✓	BLIP-CoVR	CA [41]	1	15	11.07	13.94	16.07	16.88
✓	✓	$\text{BLIP}_{\text{CoVR-ECDE}}$	CA [37]	1	15	13.03	15.90	18.62	19.83
✓	✓	TF-CoVR-Base (Stage-2 only)	MLP	8	8	15.08	18.70	21.78	22.61
✓	✓	TF-CoVR-Base (Ours)	MLP	12	12	**21.85**	**24.23**	**26.47**	**27.22**

Qualitative Results

Query Video

(**Floor Exercise**) split jump with **1 turn**

Description

Modification

show on **Balance Beam** with **0.5 turn** in side position.

Retrieval

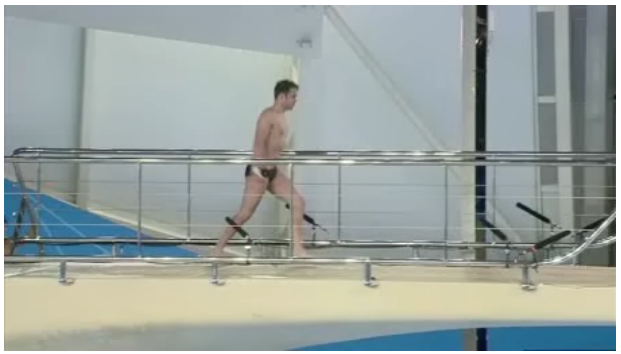
$BLIP_{CoVR\text{-}ECDE}$

TF-CoVR-Base

BLIPCoVR-ECDE fails to retrieve the correct action, which highlights the importance of accurately identifying the action class during retrieval.

Qualitative Results

Query Video



Forward, **3.5 Soms**.Pike, Entry

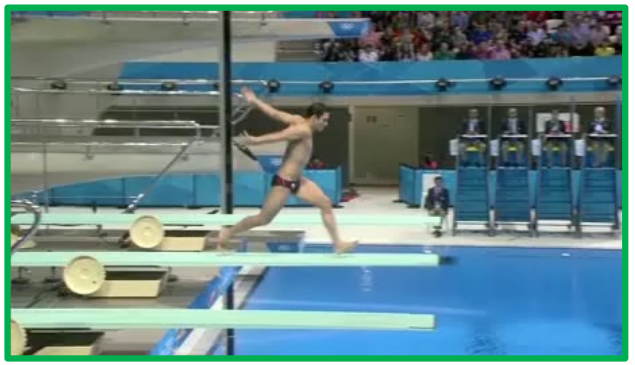
Description

Modification

show with **2.5 somersaults and 2 twists.**

Retrieval

$BLIP_{CoVR\text{-}ECDE}$



TF-CoVR-Base

BLIPCoVR-ECDE retrieves 3 twists instead of the of 2, highlighting the importance of modeling temporal action differences accurately during retrieval.

Conclusion

- A new fine-grained temporal dataset *TF-CoVR*.
  - 259 and 44 sub-actions from FineGym[1] and FineDiving[2].
  - Multiple Ground-truth instances.
  - Challenging temporal changes across the triplets.
- *TF-CoVR-Base*, a two-stage method for Composed Video Retrieval.
- Benchmark existing CoVR methods and General Multimodal Embedding (GME).

[1] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. FineGym: A Hierarchical Video Dataset for Fine-grained Action Understanding. In CVPR 2020
[2] Jinglin Xu, Yongming Rao, Xumin Yu, Guangyi Chen, Jie Zhou and Jiwen Lu. FineDiving: A Fine-grained Dataset for Procedure-aware Action Quality Assessment. In CVPR 2022.