Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

FAIR Future Artificial Intelligence Research

Bridging Creation and Comprehension: Video Generation for Video Understanding

Elisa Ricci

University of Trento & FBK

Naples, 23/09/2024

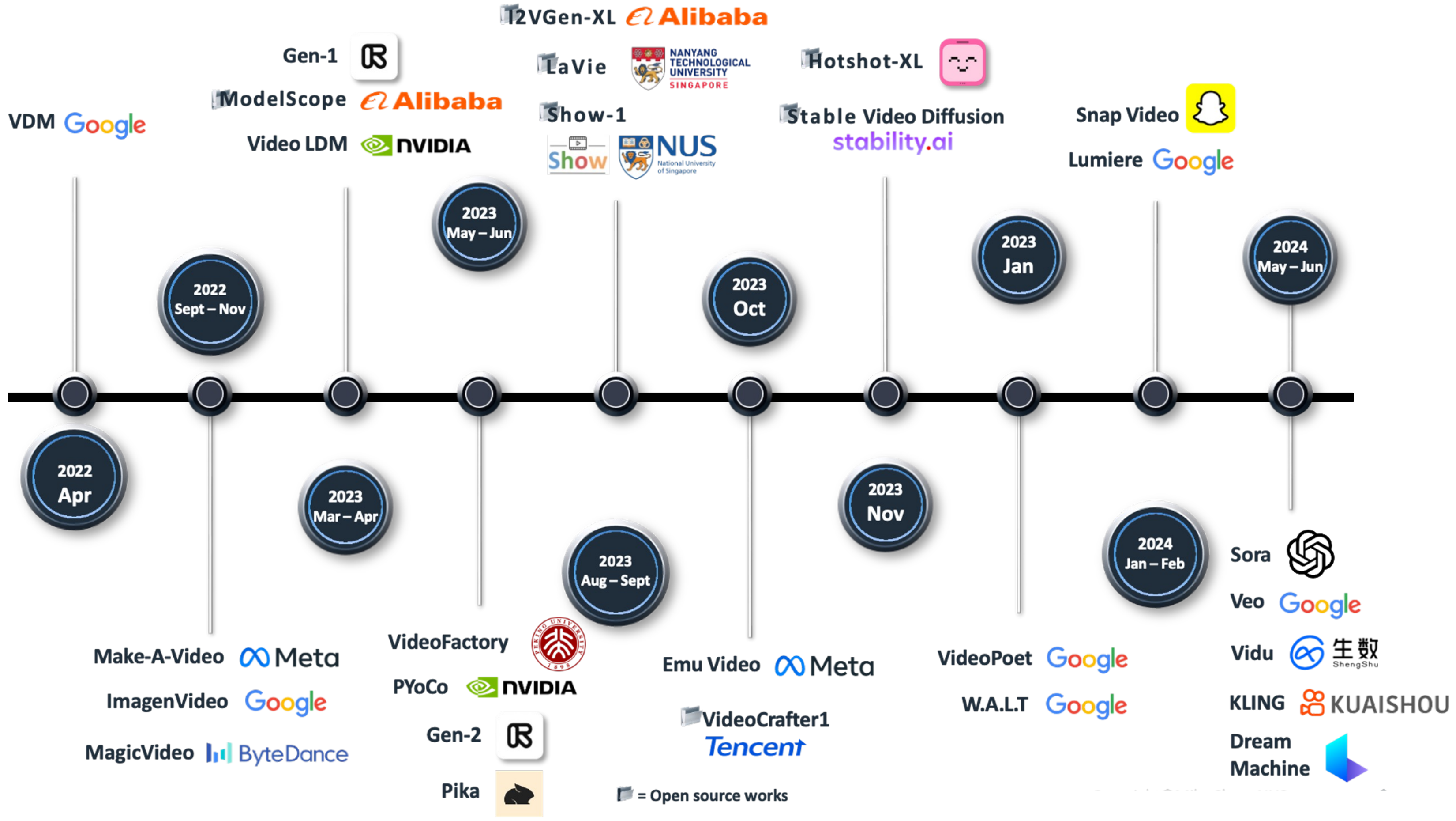
GENERAL CONFERENCE

FAIR Future Artificial Intelligence Research

23|24SET2024

NAPOLI

Video Foundation Model



2022 Apr
VDM Google

2022 Sept – Nov
Make-A-Video Meta
ImagenVideo Google
MagicVideo ByteDance

2023 Mar – Apr
Gen-1
ModelScope Alibaba
Video LDM NVIDIA

2023 May – Jun
VideoFactory
PYoCo NVIDIA
Gen-2
Pika

2023 Aug – Sept
I2VGen-XL Alibaba
LaVie NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE
Show-1 Show NUS National University of Singapore

2023 Oct
Emu Video Meta
VideoCrafter1 Tencent

2023 Nov
Hotshot-XL
Stable Video Diffusion stability.ai

2023 Jan
VideoPoet Google
W.A.L.T Google

2024 Jan – Feb
Snap Video
Lumiere Google

2024 May – Jun
Sora
Veo Google
Vidu 生数 ShengShu
KLING KUAISHOU
Dream Machine

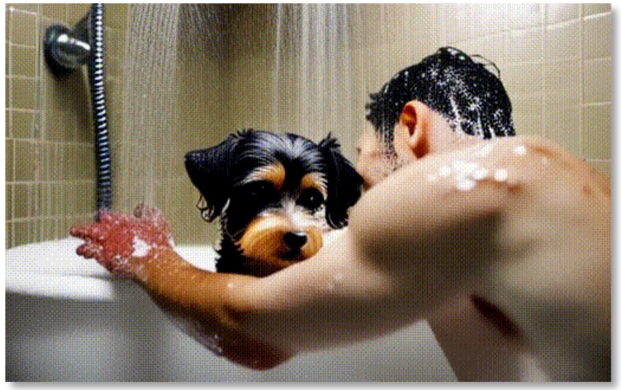
= Open source works

Text-to-Video Generation

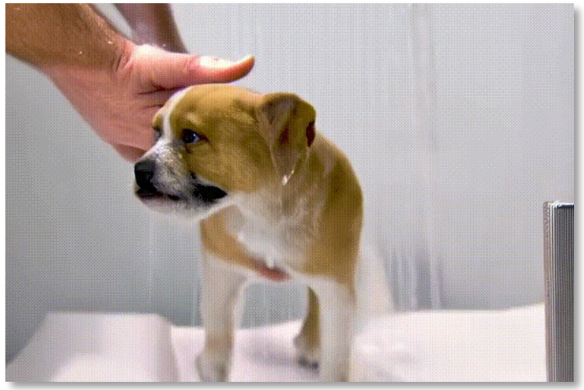
PROMPT:

“A man gives a small dog a bath in a shower.”

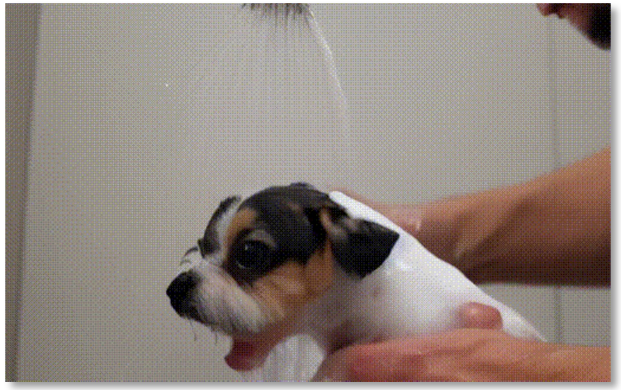
VideoCrafter2



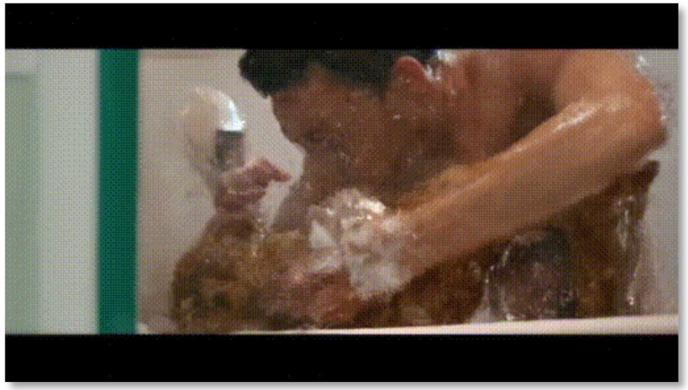
CogVideoX

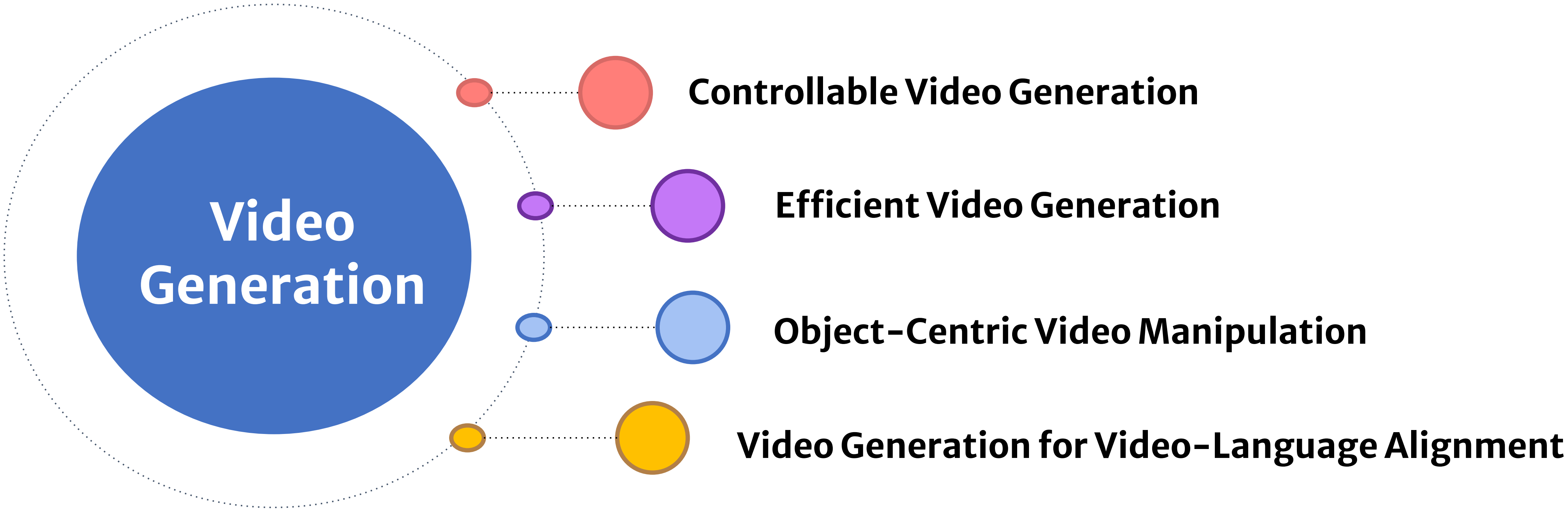


LaVie



Snap Video





Video Generation

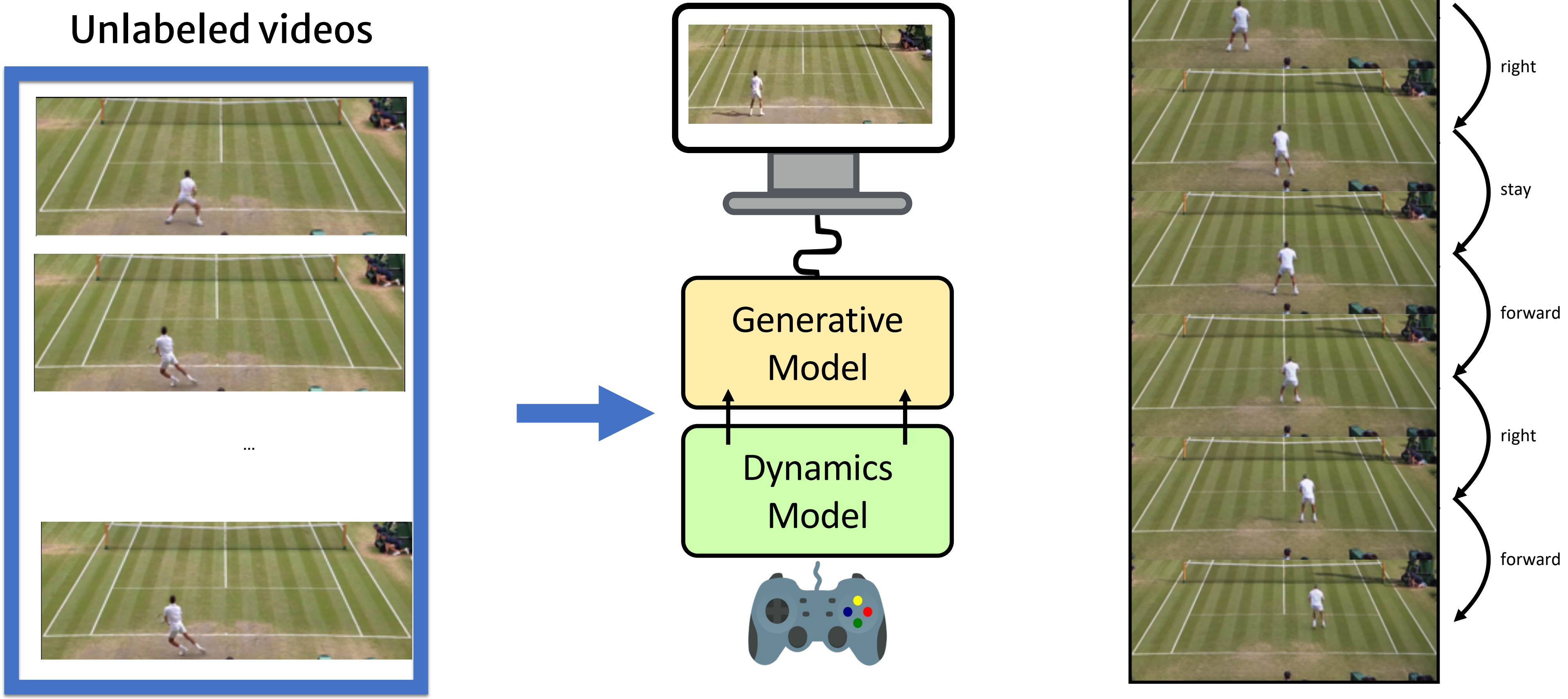
Controllable Video Generation

Efficient Video Generation

Object-Centric Video Manipulation

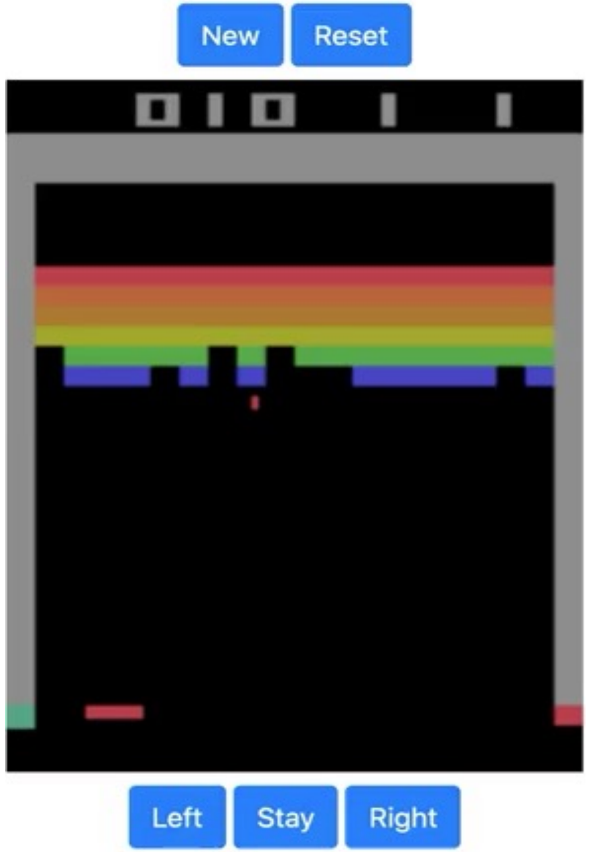
Video Generation for Video-Language Alignment

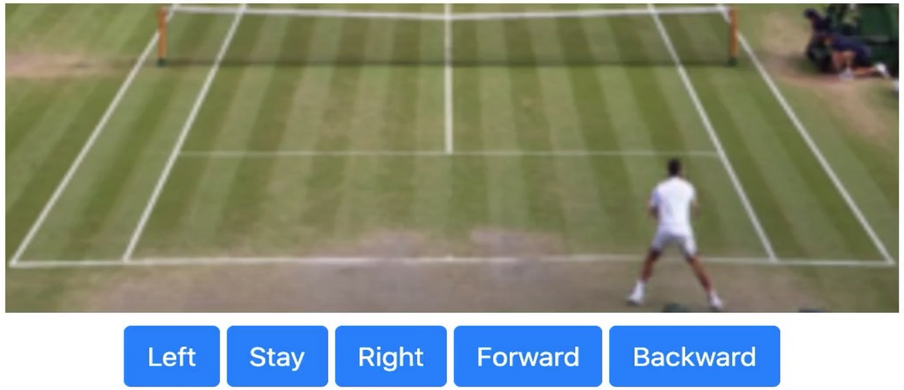
Playable Video Generation - 2021

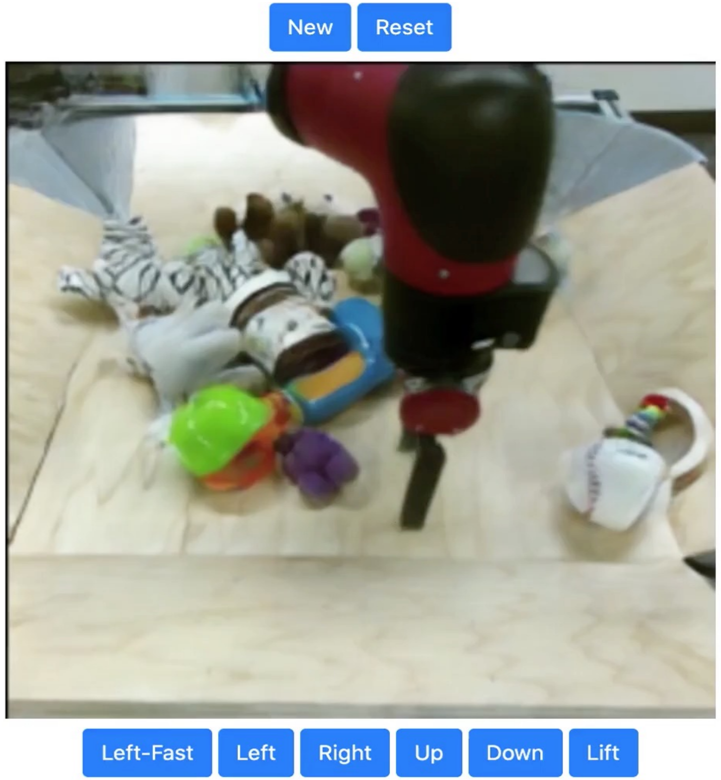


[Menapace et al. CVPR 2021]

Playable Video Generation - 2021

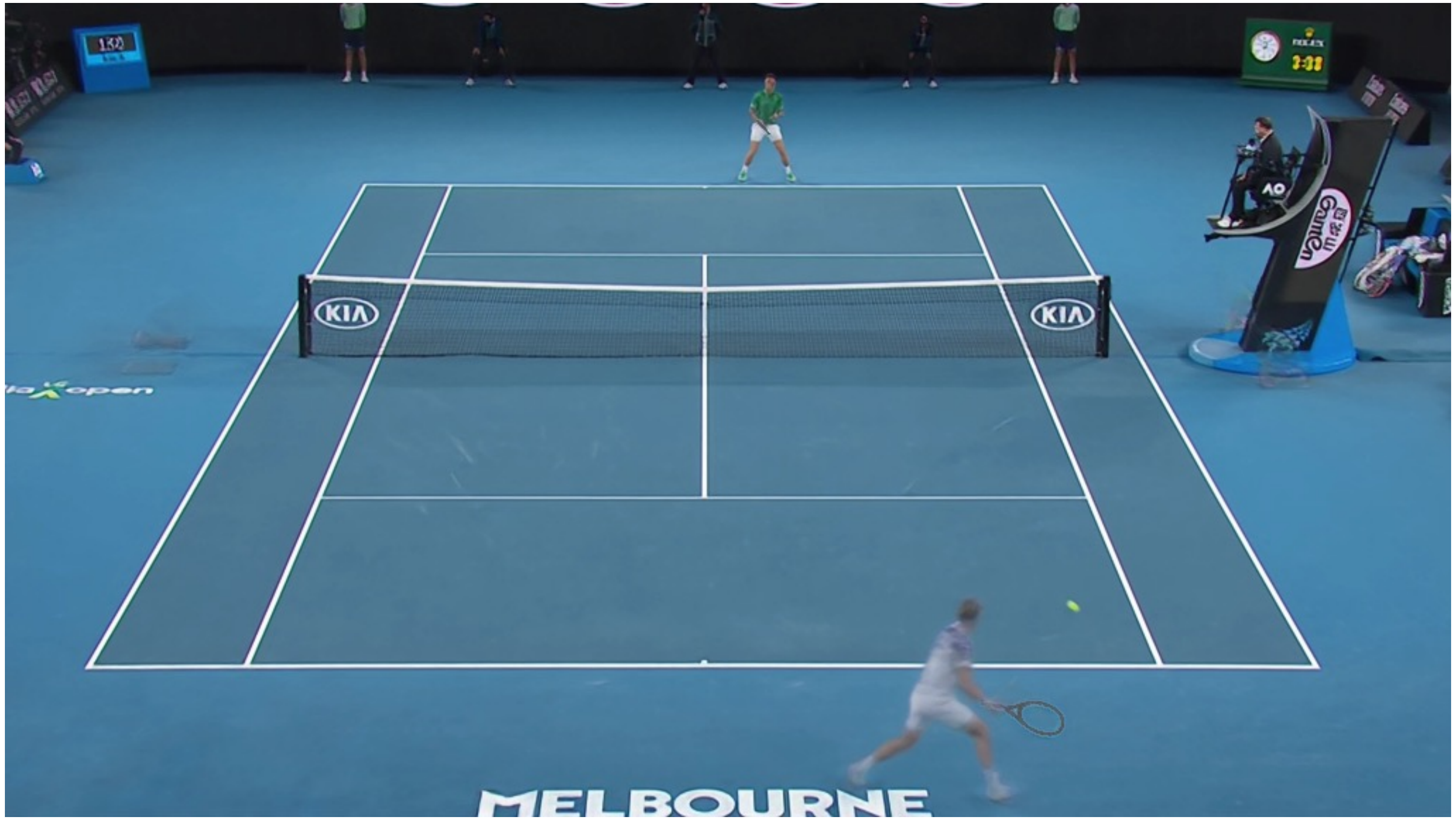






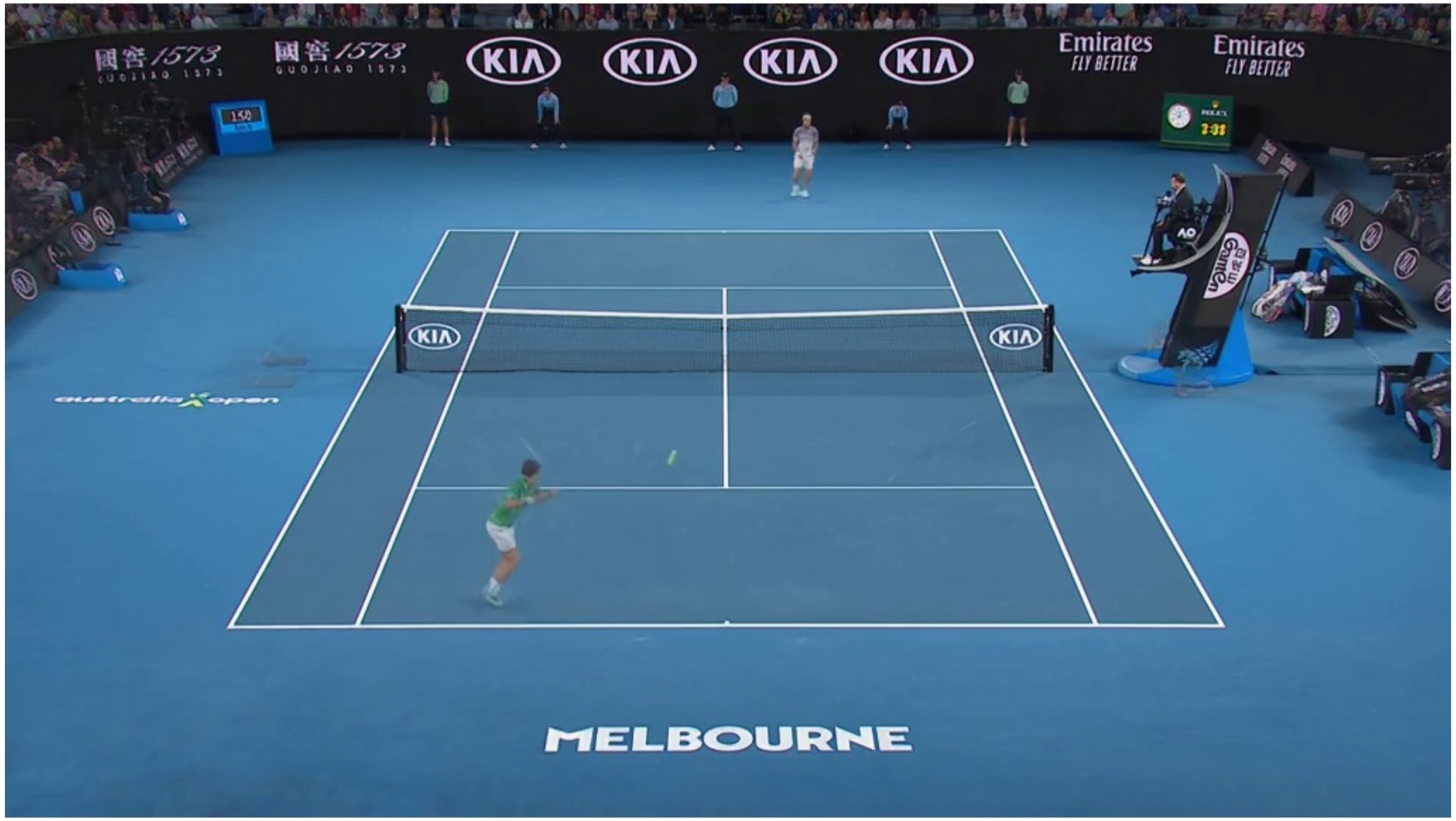
Controllable Video Synthesis

- Control the generation of (coarse) actions
- Control the camera
- Control the style



Text-Driven Video Synthesis

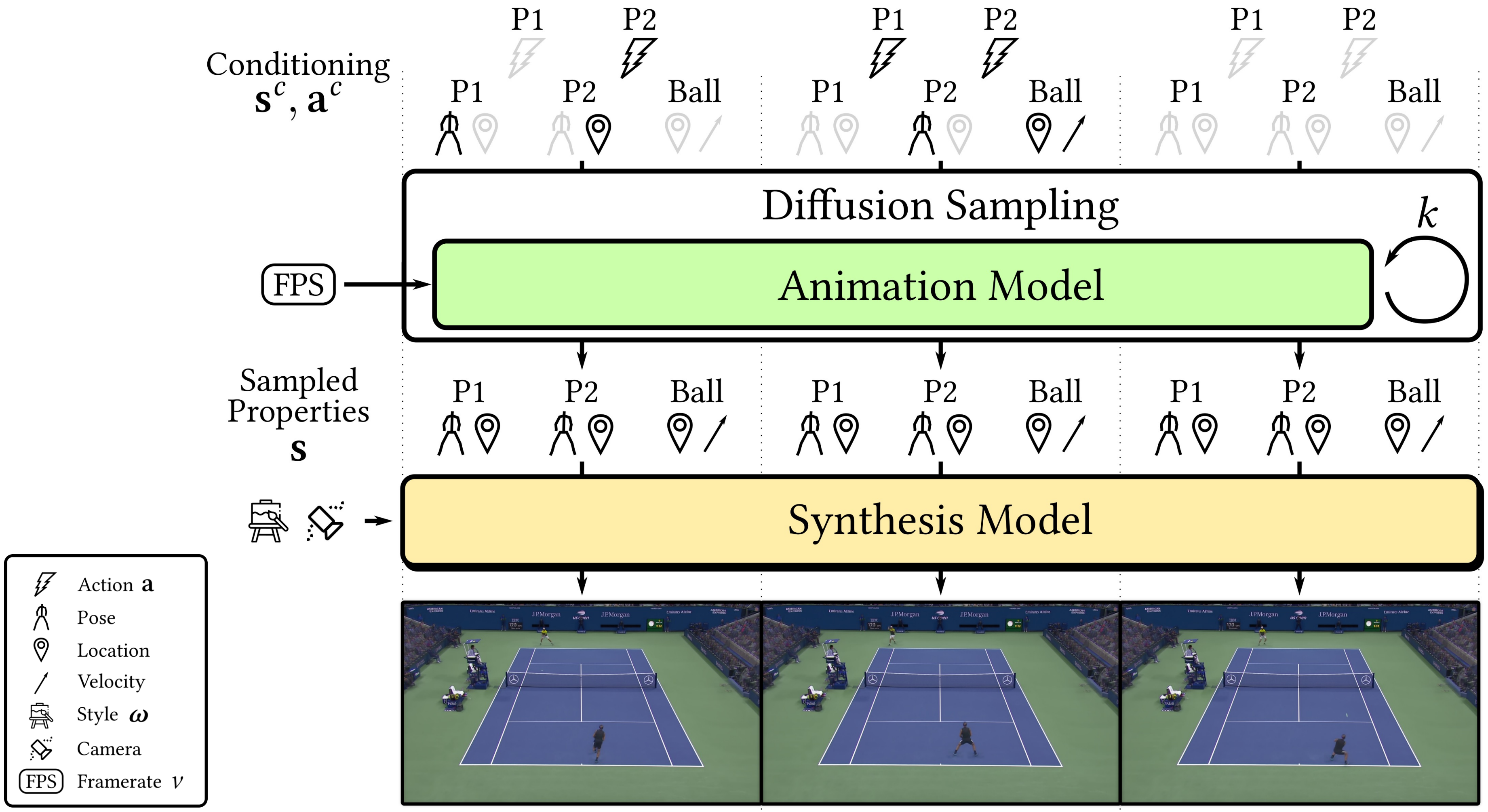
The player sidesteps to the left and stops almost in the middle behind the baseline



The player moves to the left and hits with another forehand to the right side of the field

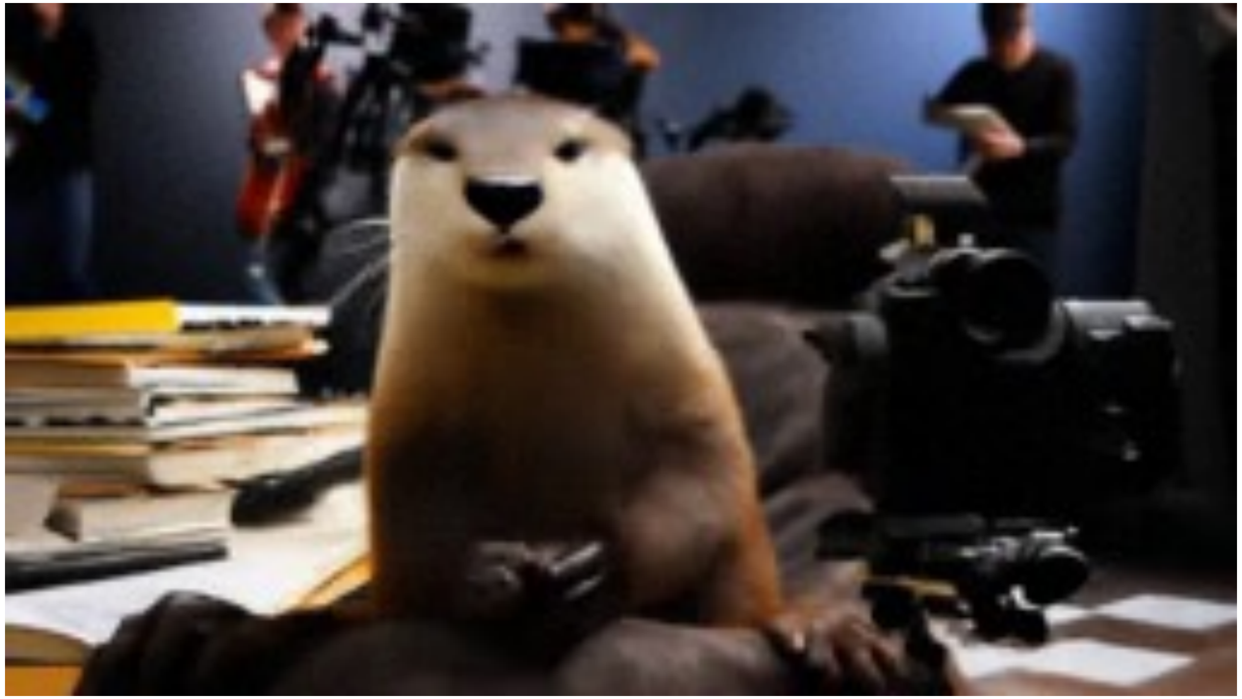
[Menapace et al. ACM TOG 2024]

Method

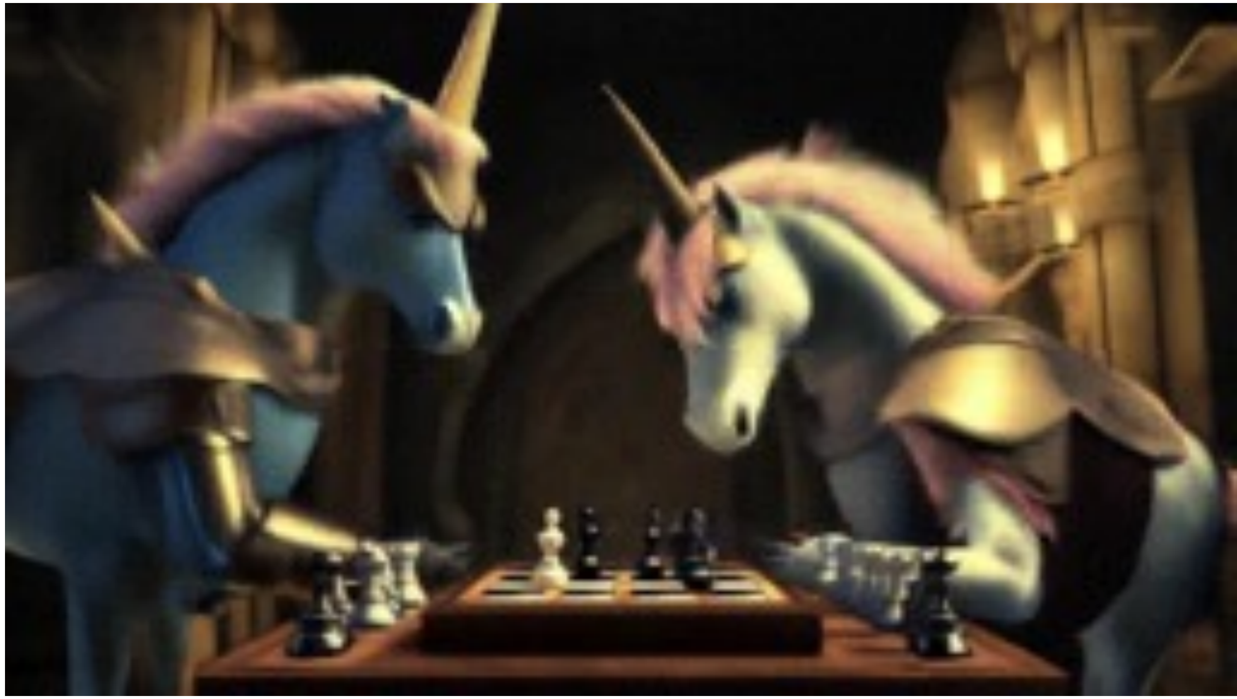


[Menapace et al. ACM TOG 2024]

Efficient Video Generation: Snap Video



“[...] a movie set where an otter serves as a film director. [...] with furrowed brows and raised paws shouting ‘Action!’”

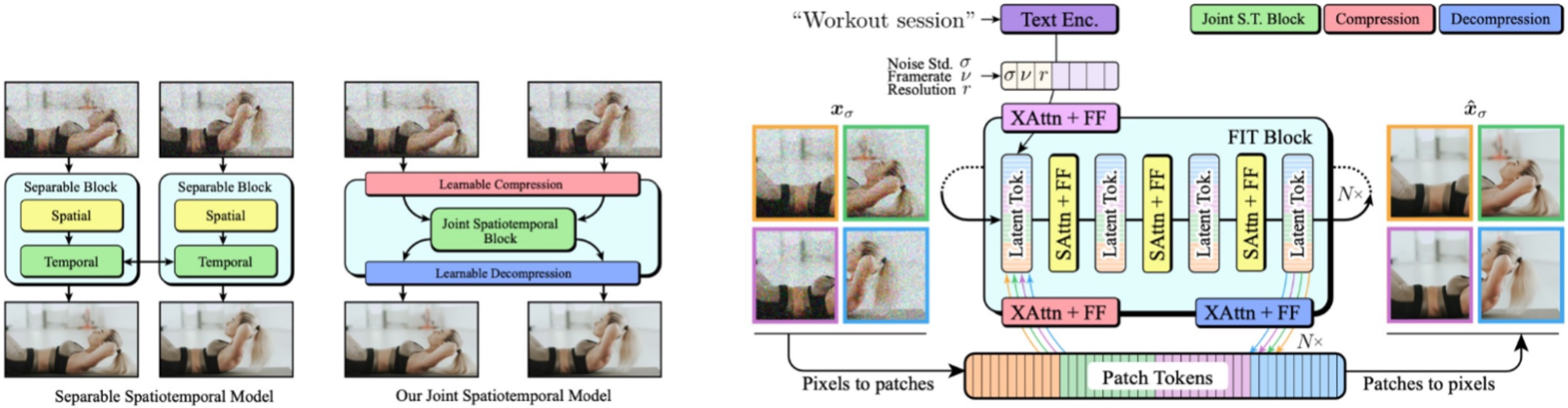


“Two unicorns in armors are playing a game of chess, in a medieval castle, high definition, photo-realistic style.”

[Menapace et al. CVPR 2024]

Efficient Video Generation: Snap Video

Transformer-based video diffusion models



(a) Computational Paradigms for Videos

(b) Snap Video FIT Architecture

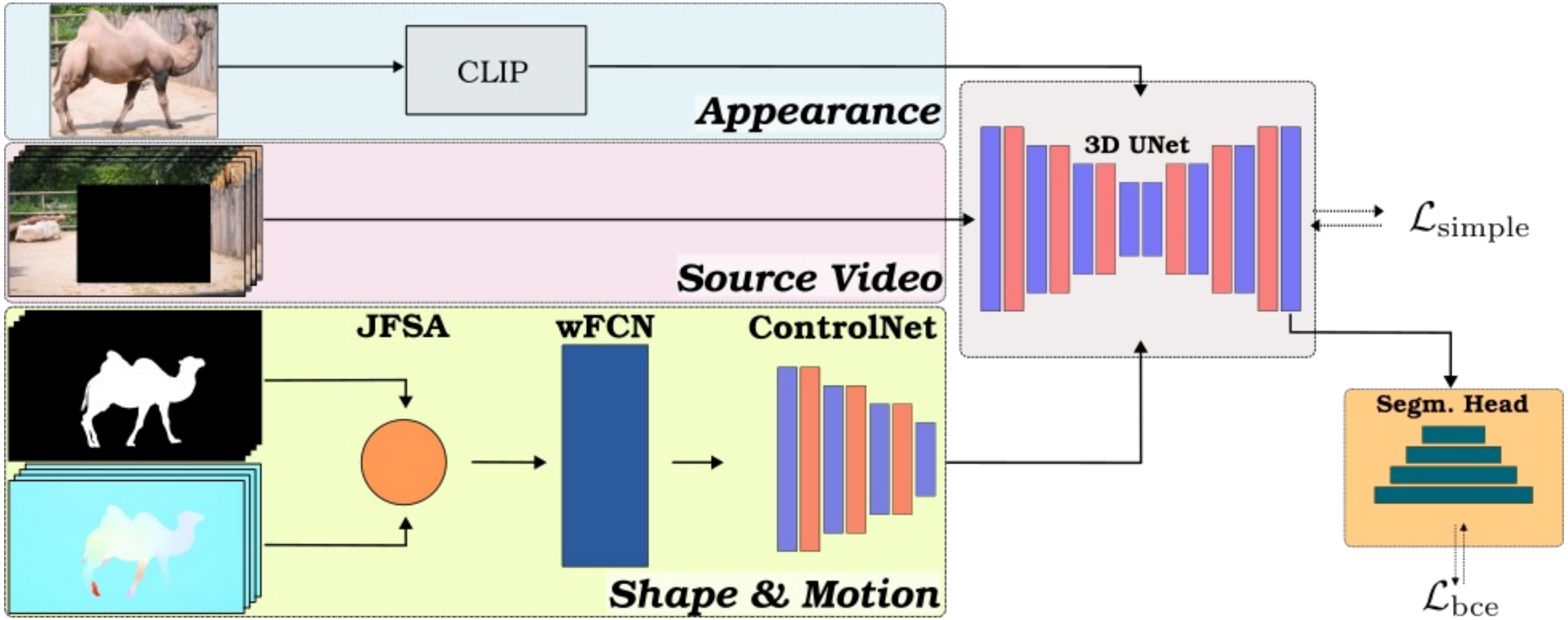
[Menapace et al. CVPR 2024]

Object-Centric Video Manipulation

We edit an object of a real video modifying the **shape** through a keyframe and the **appearance** through a driving image.

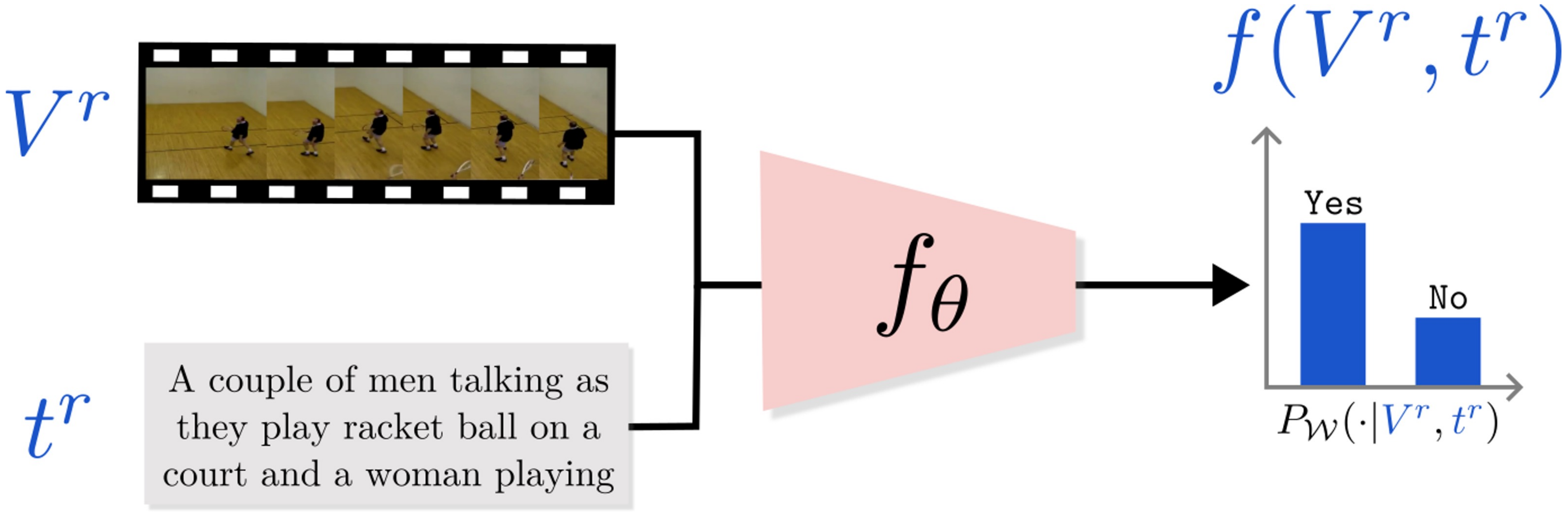
[Peruzzo et al. *under submission*]

Object-Centric Video Manipulation



[Peruzzo et al. *under submission*]

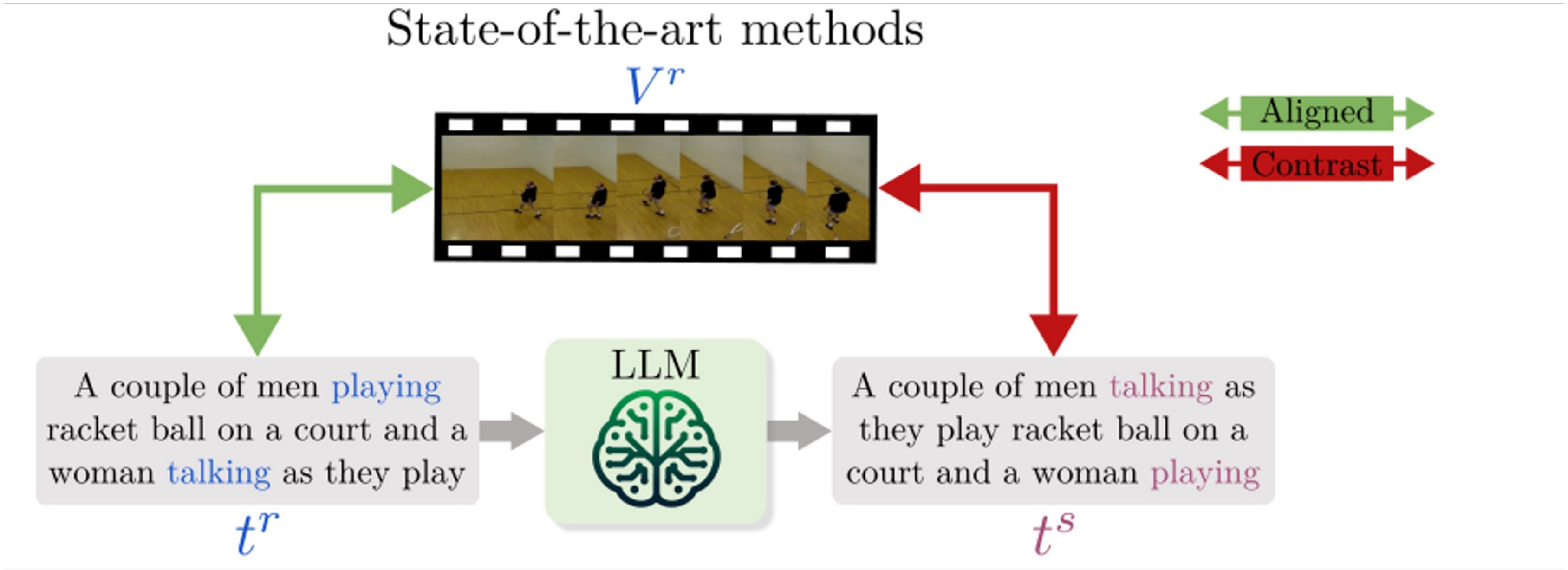
Video-Language Alignment



A couple of men talking as they play racket ball on a court and a woman playing

Lin, Zhiqiu, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. "Evaluating text-to-visual generation with image-to-text generation." In *ECCV* 2024.
Bansal, Hritik, Yonatan Bitton, Idan Szpektor, Kai-Wei Chang, and Aditya Grover. "Videocon: Robust video-language alignment via contrast captions." In *CVPR* 2024.

Video-Language Alignment



State-of-the-art methods

A couple of men playing racket ball on a court and a woman talking as they play

A couple of men talking as they play racket ball on a court and a woman playing

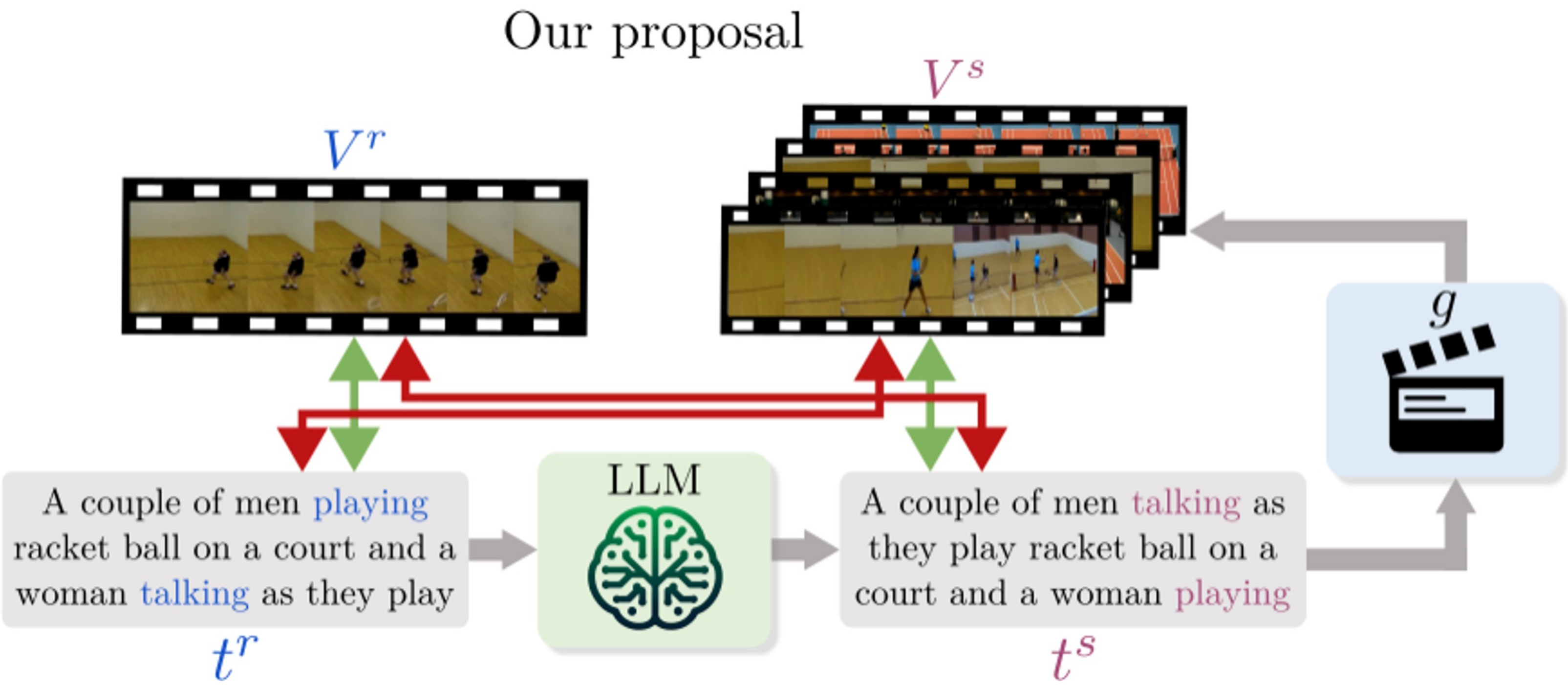
Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

FAIR Future Artificial Intelligence Research

Synthetic Videos for Video-Language Alignment

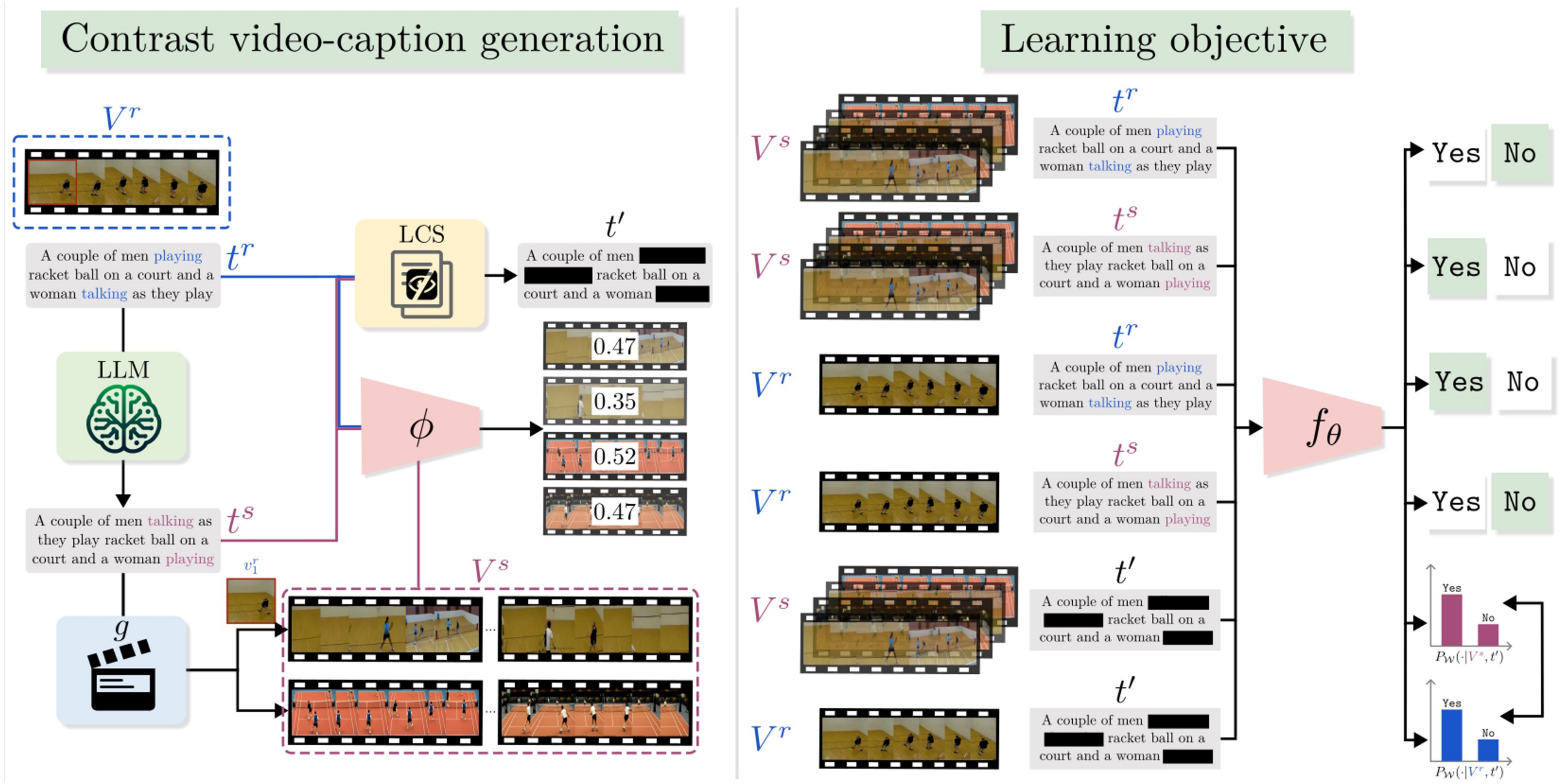


Our proposal

V^r
V^s
A couple of men playing racket ball on a court and a woman talking as they play
t^r
LLM
A couple of men talking as they play racket ball on a court and a woman playing
t^s
g

[Zanella et al. Under submission]

Method



Contrast video-caption generation

V^r

A couple of men playing racket ball on a court and a woman talking as they play t^r

LLM

A couple of men talking as they play racket ball on a court and a woman playing t^s

g

v_1^r

LCS

t'

A couple of men [redacted] racket ball on a court and a woman [redacted]

ϕ

0.47

0.35

0.52

0.47

V^s

Learning objective

V^s t^r A couple of men playing racket ball on a court and a woman talking as they play

V^s t^s A couple of men talking as they play racket ball on a court and a woman playing

V^r t^r A couple of men playing racket ball on a court and a woman talking as they play

V^r t^s A couple of men talking as they play racket ball on a court and a woman playing

V^s t' A couple of men [redacted] racket ball on a court and a woman [redacted]

V^r t' A couple of men [redacted] racket ball on a court and a woman [redacted]

f_θ

Yes No

Yes No

Yes No

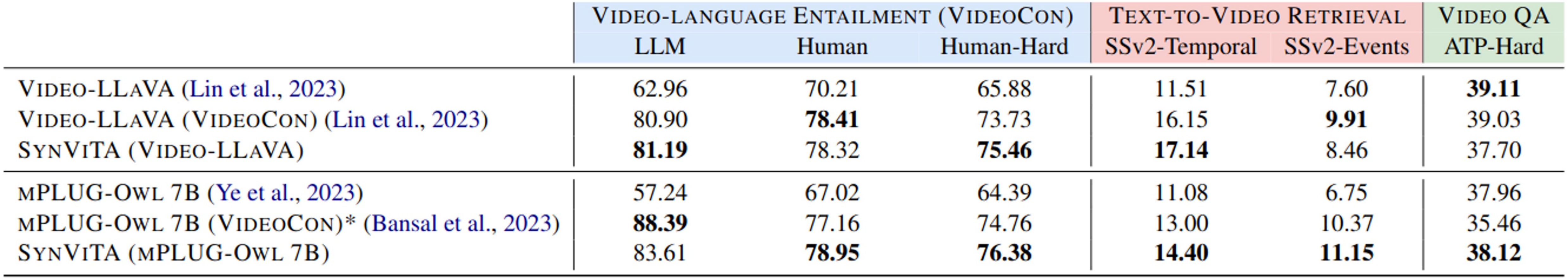
Yes No

Yes No $P_W(\cdot|V^s, t')$

Yes No $P_W(\cdot|V^r, t')$

[Zanella et al. Under submission]

Results



	Video-language Entailment (VideoCon)			Text-to-Video Retrieval		Video QA
	LLM	Human	Human-Hard	SSv2-Temporal	SSv2-Events	ATP-Hard
Video-LLaVA (Lin et al., 2023)	62.96	70.21	65.88	11.51	7.60	**39.11**
Video-LLaVA (VideoCon) (Lin et al., 2023)	80.90	**78.41**	73.73	16.15	**9.91**	39.03
SynViTA (Video-LLaVA)	**81.19**	78.32	**75.46**	**17.14**	8.46	37.70
mPLUG-Owl 7B (Ye et al., 2023)	57.24	67.02	64.39	11.08	6.75	37.96
mPLUG-Owl 7B (VideoCon)* (Bansal et al., 2023)	**88.39**	77.16	74.76	13.00	10.37	35.46
SynViTA (mPLUG-Owl 7B)	83.61	**78.95**	**76.38**	**14.40**	**11.15**	**38.12**

[Zanella et al. Under submission]

Conclusions

- Numerous models for video generation (with many more on the horizon)
- Synthetic videos offer immense value, also for improving video understanding systems.
- Current challenges: misalignment between generated videos and prompts, scalability issues, appearance gap between synthetic and real videos.
- Wide range of applications: video surveillance, multimedia analysis, robotics, etc.

Thank you!

GENERAL CONFERENCE

FAIR Future Artificial Intelligence Research

23|24SET2024

NAPOLI