ECAI2025

Unlocking the Potential of mLLMs: Enhancing Video-Text Retrieval through Caption Supplementation andConical Embedding Optimization

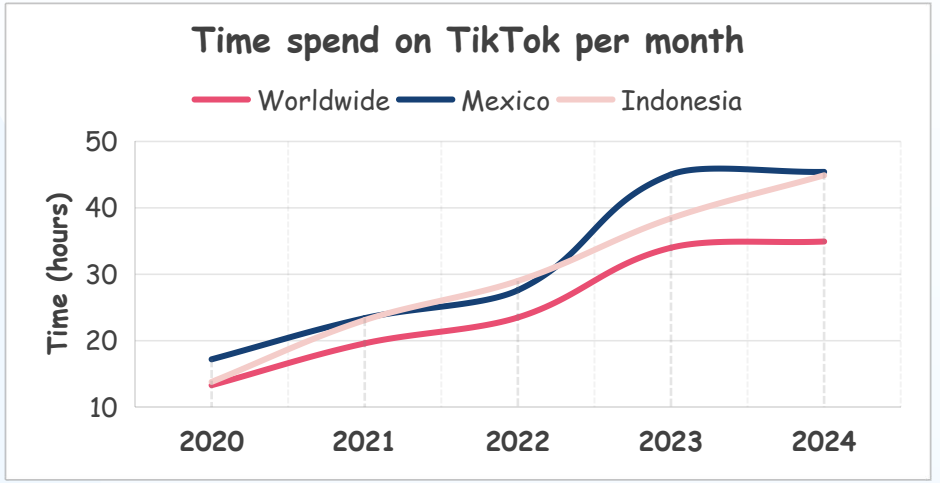
Baoyao Yang, Junxiang Chen and Wenbin Yao

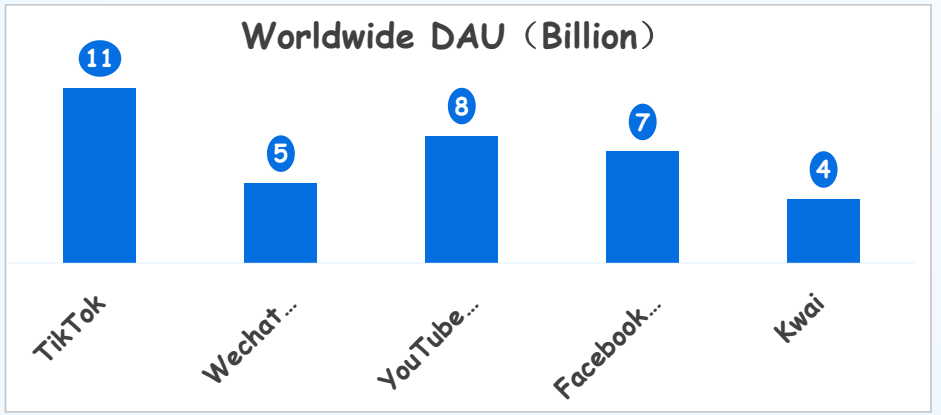
廣東工業大學
Guangdong University of Technology

视频号

OCTOBER 17, 2025

Background

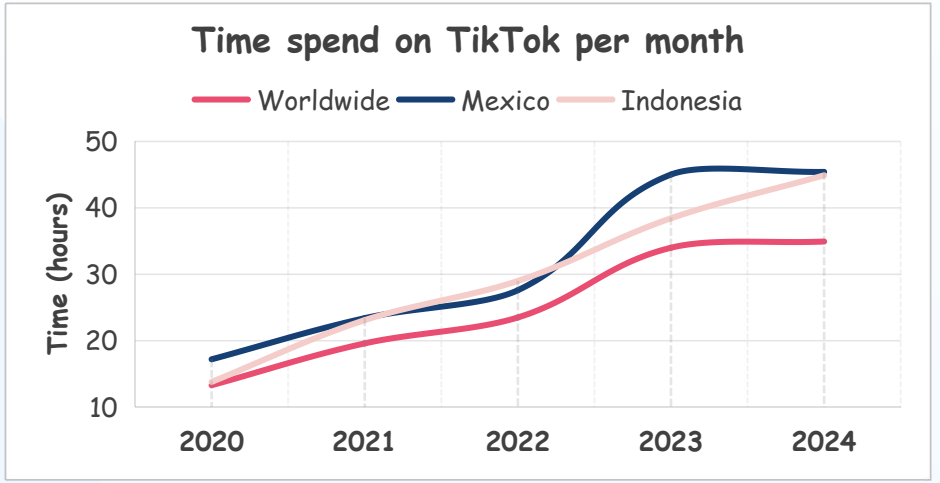


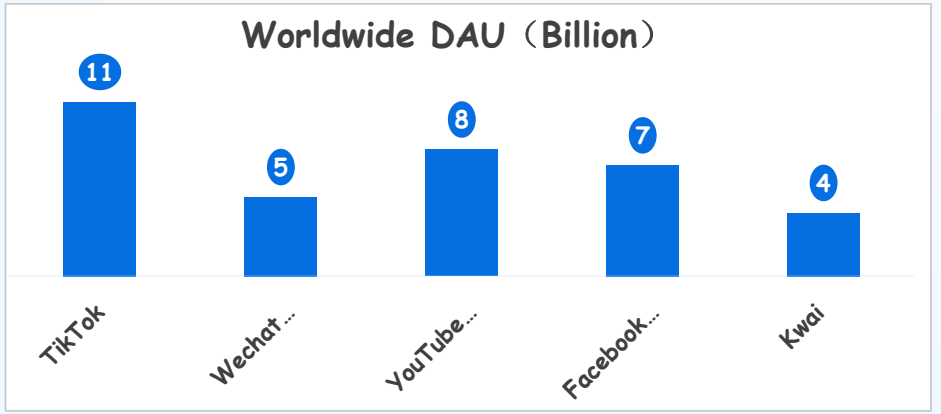


- Recommendation
- Quality Assessment
- Content Search
- Hotpot Summarize
- ...

Short videos' popularity on social media has driven sharp demand for video-text understanding

Background



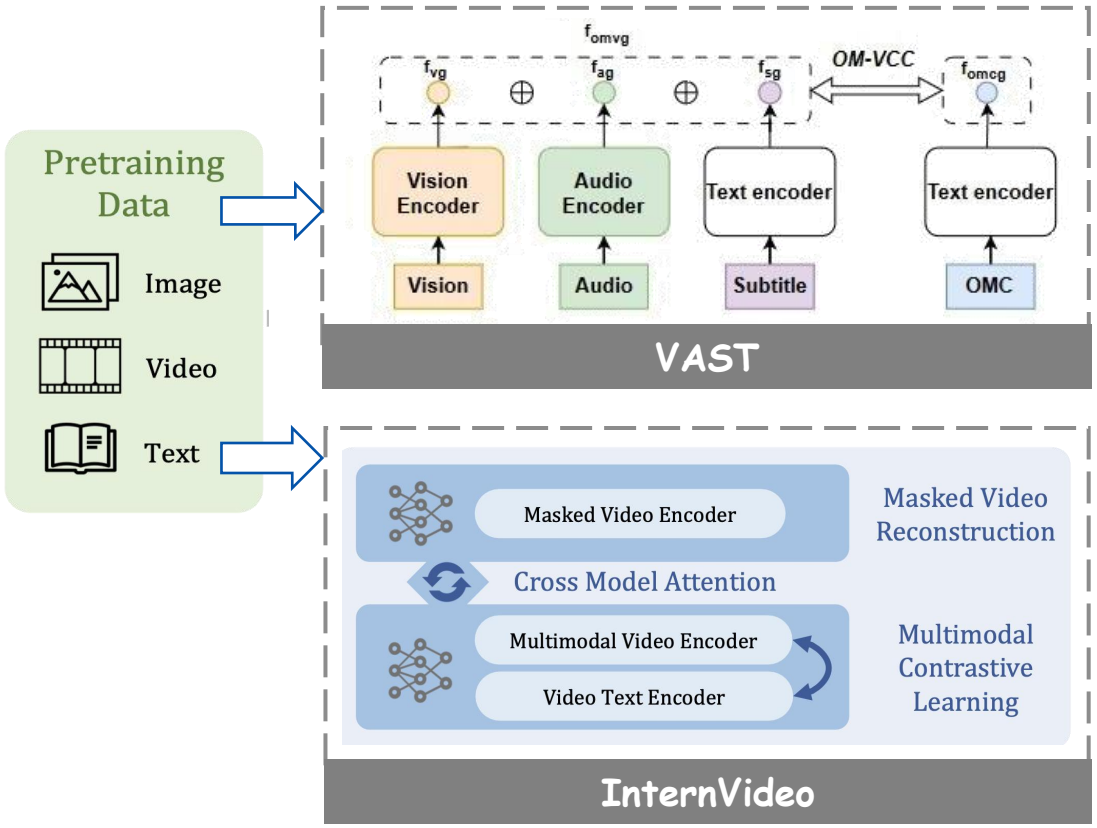


- Recommendation
- Quality Assessment
- Content Search
- Hotpot Summarize
- ⋮

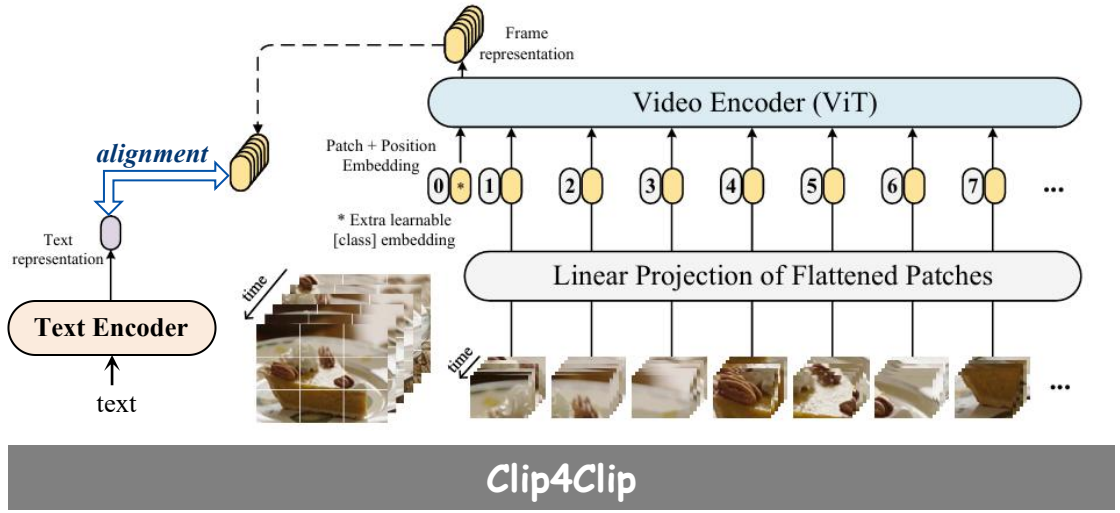
Core research → **Video-Text Retrieval**

Mainstream Solution & Limitation

Pretraining multimodal foundation models

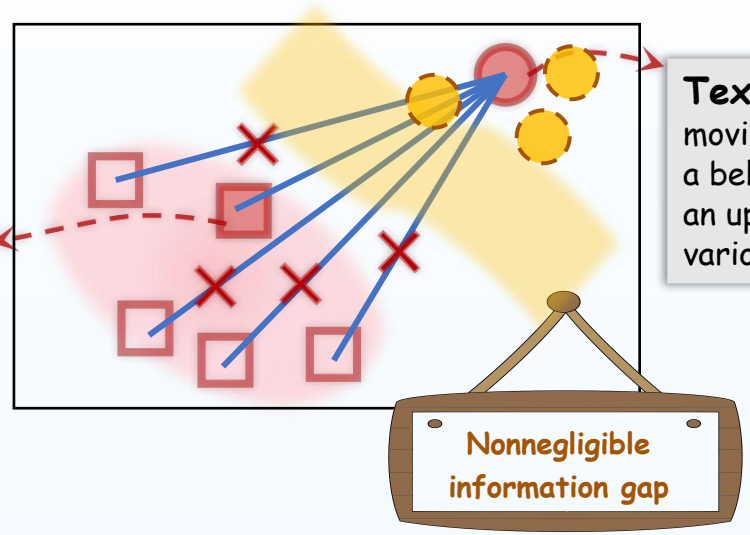


Adapt image-text model to videos



Heavy workload on data collection

Key Research Issue in VTR

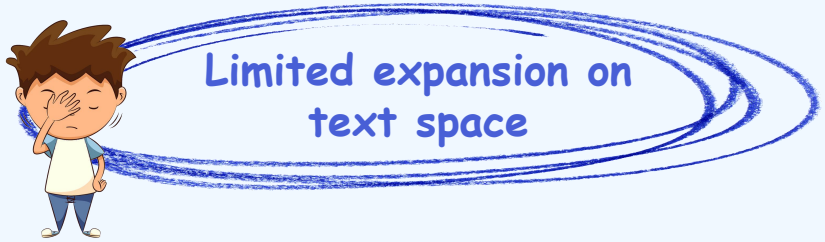


Text
movie producers are giving a behind the scenes look at an upcoming movie in various settings

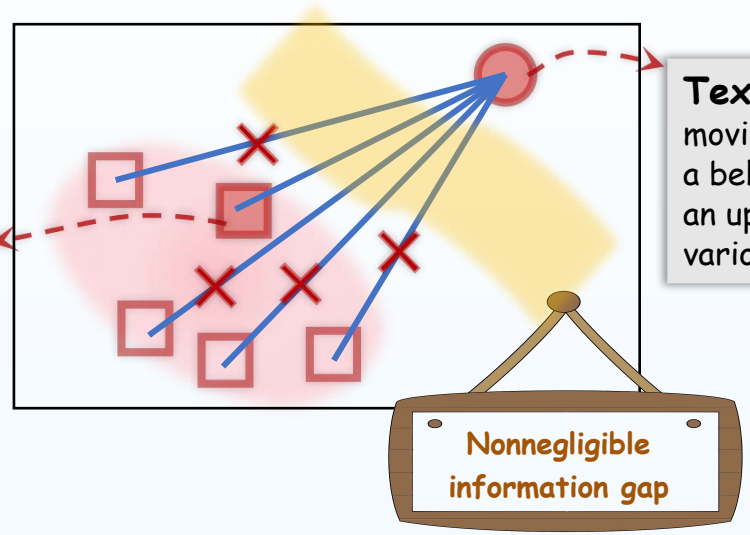
Single frame discriptions
Overly simplistic and one-sided

Text rewrite — *effective approach eliminating data collection*

- **CAD** [ECCV 2024]: Rewrite action descriptions using human-machine collaborative systems
- **Cap4Video++** [TPAMI 2024]: Repharsing text using CLIP+text-based LLM



Key Research Issue in VTR



Text
movie producers are giving a behind the scenes look at an upcoming movie in various settings

Nonnegligible information gap

Single frame discriptions
Overly simplistic and one-sided

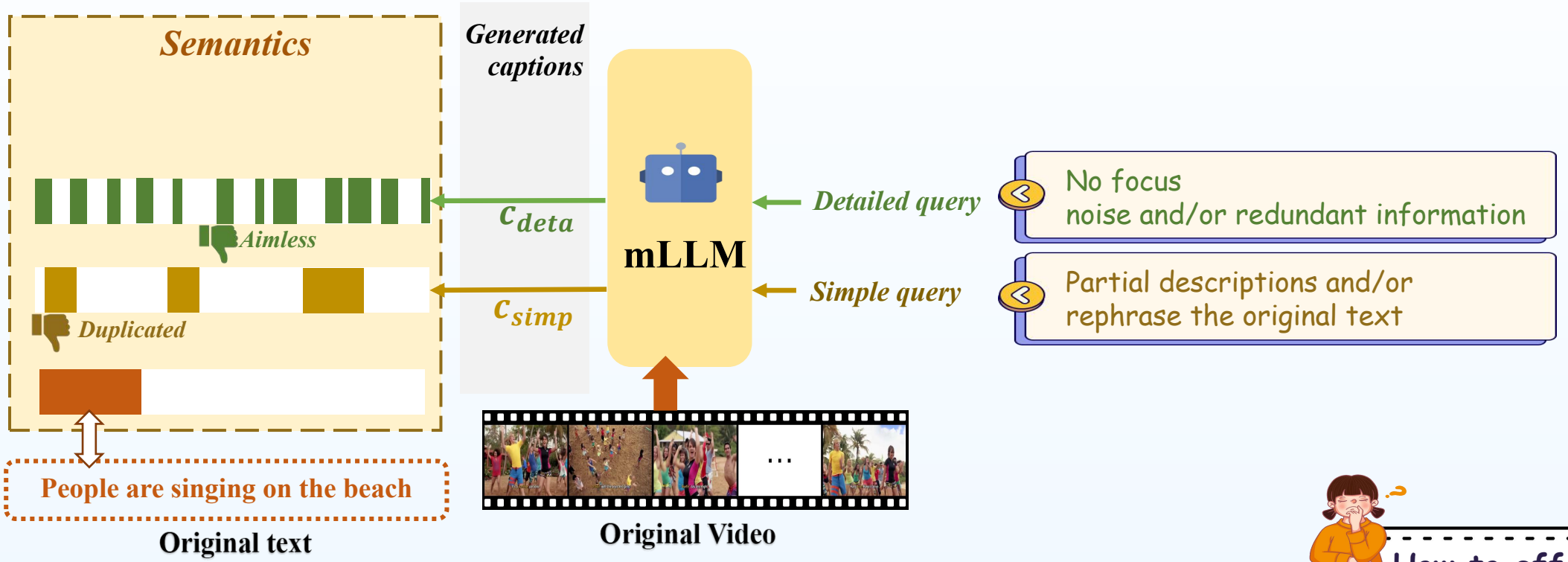
mLLM trianed by large scale multimodal data

Reveal vast knowledge encapsulated in mLLM to bridge modality gap

How to effectively leverage mLLMs to enhance VTR ?

Generation **Utilization**

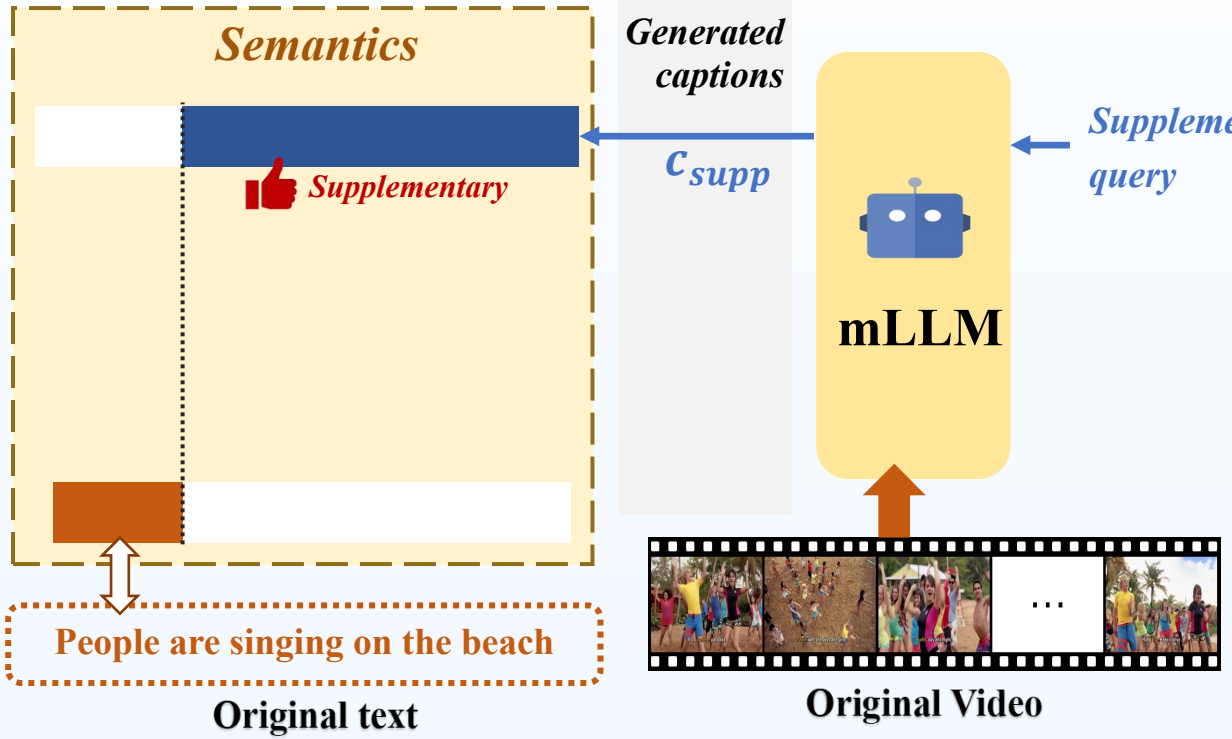
Key Research Issue in VTR



How to effectively leverage mLLMs to enhance VTR ?

Generation

Our Proposal (mLLM-supplemental Captions)



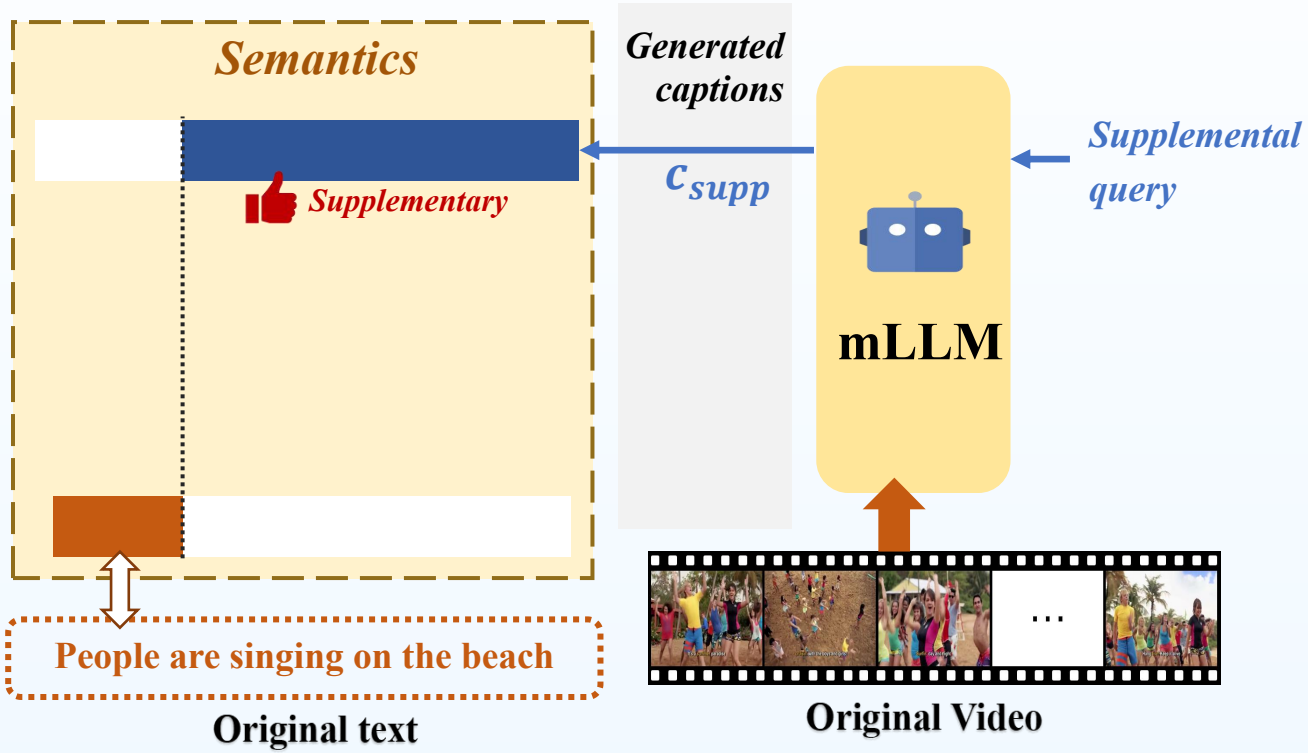
Key Strategies

- Uncovered Information
- Detailed Expressions
- Avoid Redundancy
- Co-consideration on video

How to effectively leverage mLLMs to enhance VTR ?

Generation

Our Proposal (mLLM-supplemental Captions)



Some examples of mLLM-supplemental captions

C_{org}: a person is cooking ← original text

C_{sup}: The video showcases a cooking process … is seen adding cornstarch to a pot of liquid. The person then stirs the mixture … a chef or home cook sharing … suggest a focus on healthy or gluten-free cooking...

new video content in supplemental captions

C_{org}: a boy is performing for judges

C_{sup}: A young child takes center stage, … the crowd erupts into applause, ... The child's face beams with pride and joy, …. the child has a natural talent for singing, and his performance has touched the hearts of the viewers…

C_{org}: some models are walking on a cat walk

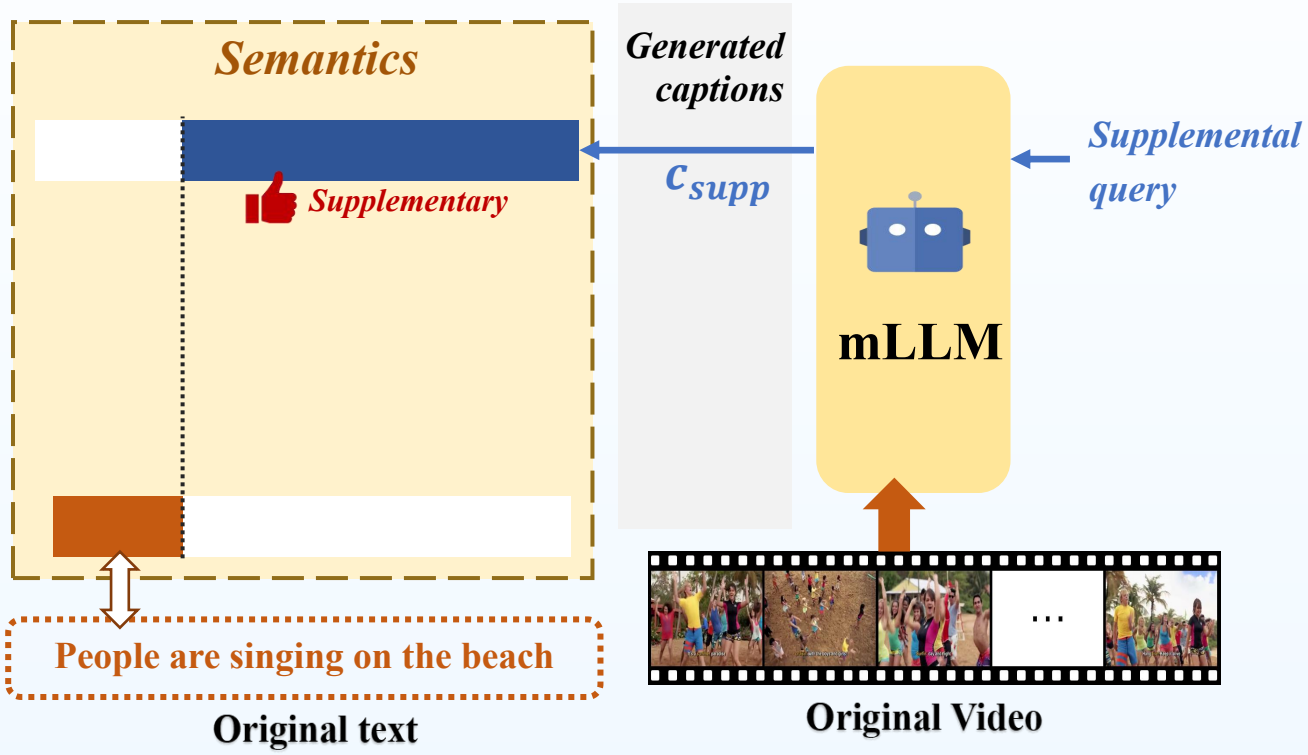
C_{sup}: The models on the runway are dressed in stunning red dresses, each with unique designs and patterns. …., the audience watches in awe, … feel drawn to the energy and excitement of the show ….

leverage VTR ?

Generation

Promote cross-modal information balance by revealing the vast knowledge encapsulated in large models

Our Proposal (mLLM-supplemental Captions)



How to effectively leverage mLLMs to enhance VTR ?

Generation

Utilization

Methodology (Conical Embedding Optimization, CEO)

- Consider text embeddings as out-of-space points, performing video-text alignment through compressing the conical-like representation space

CEO Dual Objectives

Elasticity Preservation

- Random transformation to enrich visual diversity
- InfoNCE loss: Constrict visual space to preserve representativeness

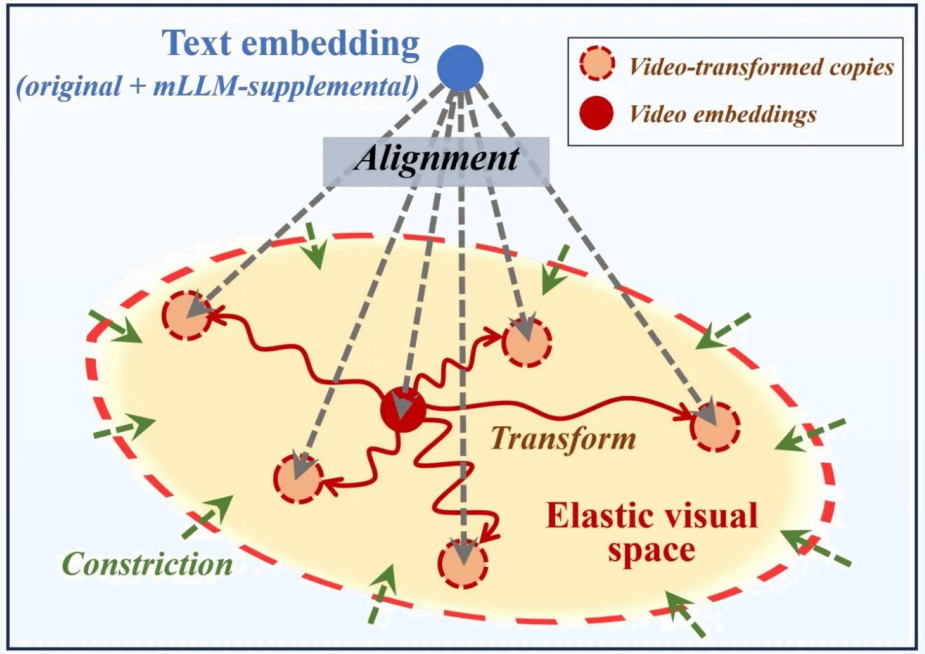
$$L_{va} = -\mathbb{E}[log\frac{exp(\tau_{va} \cdot sim(v^i_{org}, v^i_{aug}))}{\sum_{i \neq j} exp(\tau_{va} \cdot sim(v^i_{org}, v^j_{aug}))}]$$

Cross-modal Alignment on Conical-like space

- Contrastive loss: Align to both video and its transformed copy

$$L_{tv} = -\mathbb{E}[log\frac{exp(\tau_{tv} \cdot sim(t^i_{emb}, v^i_{org}))}{\sum_{i \neq j} exp(\tau_{tv} \cdot sim(t^i_{emb}, v^j_{org}))}]$$

$$L_{ta} = -\mathbb{E}[log\frac{exp(\tau_{ta} \cdot sim(t^i_{emb}, v^i_{aug}))}{\sum_{i \neq j} exp(\tau_{ta} \cdot sim(t^i_{emb}, v^j_{aug}))}]$$



Experiments: Setup

Datasets

- **MSR-VTT:** 9,000 train, 1,000 test
- **MSVD:** 1,970 videos, 80K captions
  Train/val/test split: 1,200/100/670
- **DiDeMo:** 8,395 train, 1,065 val, 1,004 test

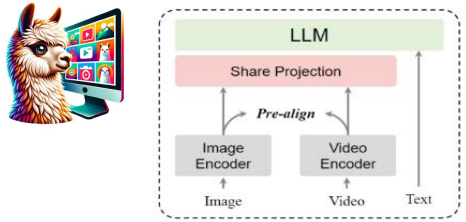
Preprocessing

Text: CLIP tokenizer
sentences split into max 70 tokens

Video: 10 frames sampled
randomly rotated/flipped for augmentation

mLLM-supplemental caption generatin

mLLM selection: Video-LLava



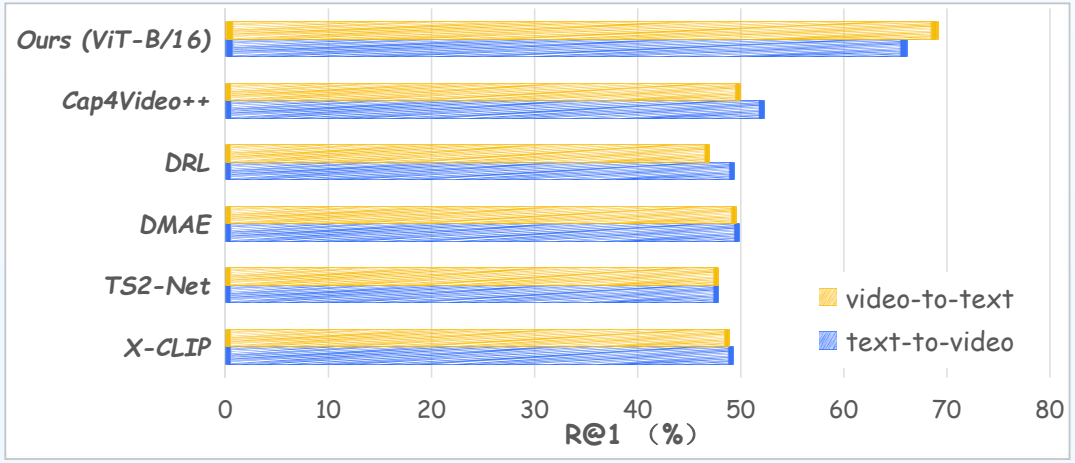
Data source: video-text pairs of each dataset

Prompts: 3 complexity levels

	Complexity	Details
Prompt A	Detailed	Generate new caption introducing additional layers of interpretation not evident in the original.
Prompt B	Medium	Generate brief captions including details and describing it succinctly.
Prompt C	Simple	Describe this video.

Experiments: SOTA Comparison (MSR-VTT)

Method	Extra data	text-to-video				video-to-text			
		R@1↑	R@5↑	R@10↑	MeanR↓	R@1↑	R@5↑	R@10↑	MeanR↓
ViT-B/32									
DRL [31]	✓	47.5	73.8	83.6	13.3	46.3	72.7	82.5	9.5
TS2-Net [18]	✗	47.2	73.7	83.1	13.1	44.8	74.3	84.0	9.3
Clip4Clip [19]	✓	44.5	71.4	81.6	15.3	42.7	70.9	80.6	11.6
X-CLIP [21]	✗	46.1	73.0	83.1	13.2	46.8	73.3	84.0	9.1
CLIP-VIP [38]	✓	55.9	77.0	86.8	–	–	–	–	–
DMAE [10]	✗	46.9	74.6	84.2	12.8	46.2	73.7	84.2	8.8
Cap4Video++ [33]	✗	50.3	75.8	85.4	12.0	47.9	74.9	85.1	8.3
TeachClip [27]	✗	46.8	74.3	82.6	–	–	–	–	–
Ours	✗	**64.2**	**90.0**	**94.3**	**3.3**	**67.3**	**90.0**	**94.5**	**3.1**
ViT-B/16									
DRL [31]	✓	49.4	76.4	84.2	13.2	47.0	77.1	84.4	9.2
TS2-Net [18]	✗	47.8	76.8	85.2	13.7	47.8	76.0	84.6	8.5
Clip4Clip [19]	✓	46.4	72.1	82.0	14.7	45.4	73.4	82.4	10.7
X-CLIP [21]	✗	49.3	75.8	84.8	12.2	48.9	76.8	84.5	8.1
CLIP-VIP [38]	✓	57.7	80.5	88.2	–	–	–	–	–
DMAE [10]	✗	49.9	75.8	85.5	12.5	49.6	76.3	85.0	8.5
Cap4Video++ [33]	✗	52.3	76.8	85.8	11.5	50.0	75.9	86.0	7.8
TeachClip [27]	✗	48.0	75.9	83.5	–	–	–	–	–
Ours	✗	**66.2**	**91.1**	**94.6**	**2.9**	**69.2**	**89.6**	**94.4**	**2.8**



Main Findings

Superior pecrformance

66.2% R@1 for t2v and 69.2% R@1 for v2t retrieval

Significant improvement

Approximately 8% increase over existing methods

Competitive advantage

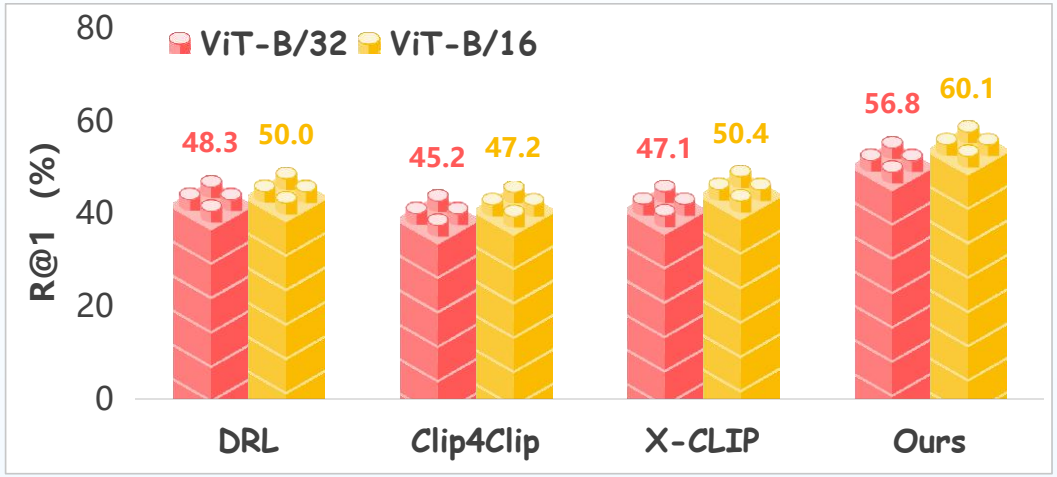
Outperformed methods using additional data

Comprehensive evaluation

Consistent improvements across all metrics

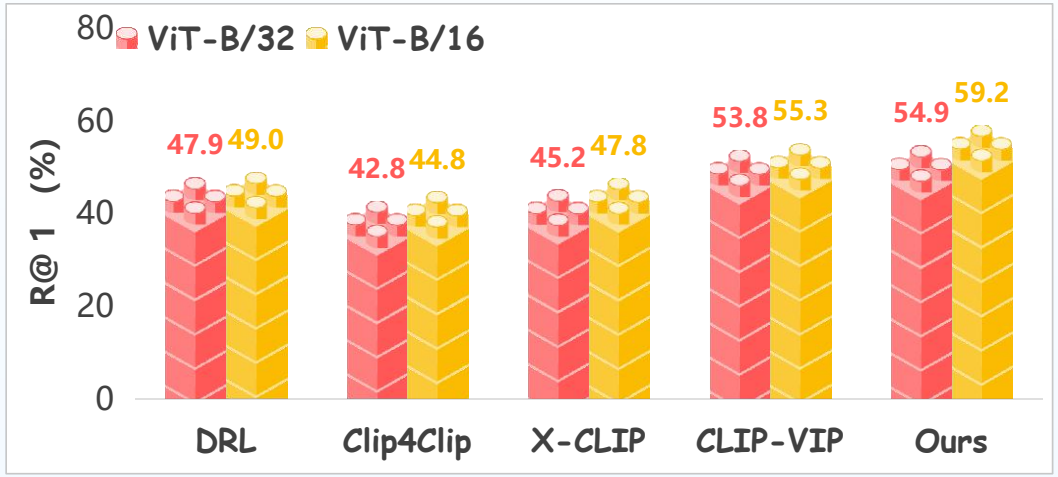
Experiments: SOTA Comparison (more)

MSVD Dataset (text-to-video)



ViT-B/32: R@1: 56.8%, R@5: 81.8%, R@10: 89.5%
ViT-B/16: R@1: 60.1%, R@5: 82.7%, R@10: 89.6%

DiDeMo Dataset (text-to-video)



ViT-B/32: R@1: 54.9%, R@5: 83.7%, R@10: 91.2%
ViT-B/16: R@1: 59.2%, R@5: 85.4%, R@10: 91.8%

- Consistent improvements over baseline methods
- ViT-B/16 configuration shows best performance
- Top-10 retrieval accuracy exceeds 90% on both datasets

Consistent performance gains across MSVD and DiDeMo datasets ⇨ Robustness across different domains

Experiments: Qualitative Results

Text-to-Video Retrieval (Vit-B/16)

- ✓ *Successfully* retrieves the corresponding real video in the first position
- ✓ Second and third hits are also highly relevant

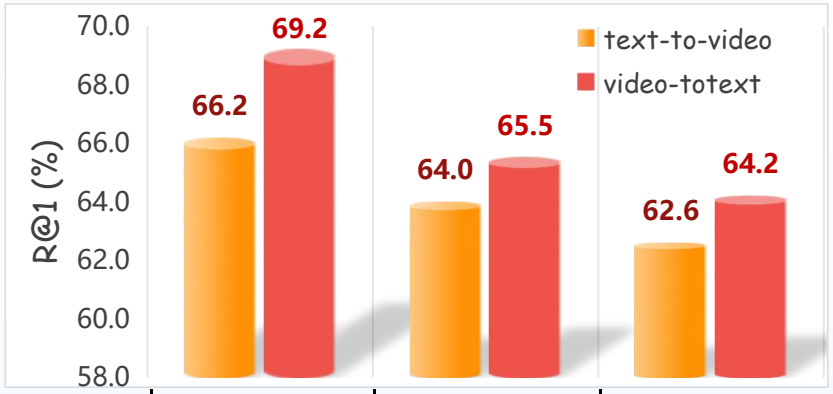
Video-to-Text Retrieval (Vit-B/16)

R1	**Base**: fireworks are being lit and exploding in a night sky **Ours**: the lighting work is going on the building	**Base**: an animated girl talks to a baby and plays with it **Ours**: cartoon play for kids	**Base**: a man is singing **Ours**: a man is singing on stage to a huge audience he is holding a microphone
R2	**Base**: women stand on a platform suspended high above the city **Ours**: city limits photograph taken from high point in day time	**Base**: cartoon girl is talking **Ours**: a cartoon shows two dogs talking to a bird	**Base**: a man is singing on stage to a huge audience he is holding a microphone **Ours**: tom jones performing live on a television show
R3	**Base**: the lighting work is going on the building **Ours**: fireworks are being lit and exploding in a night sky	**Base**: animated comic scene of guy cutting up food for dinner **Ours**: cartoon show for kids	**Base**: tom jones performing live on a television show **Ours**: people enjoy the performance of singer

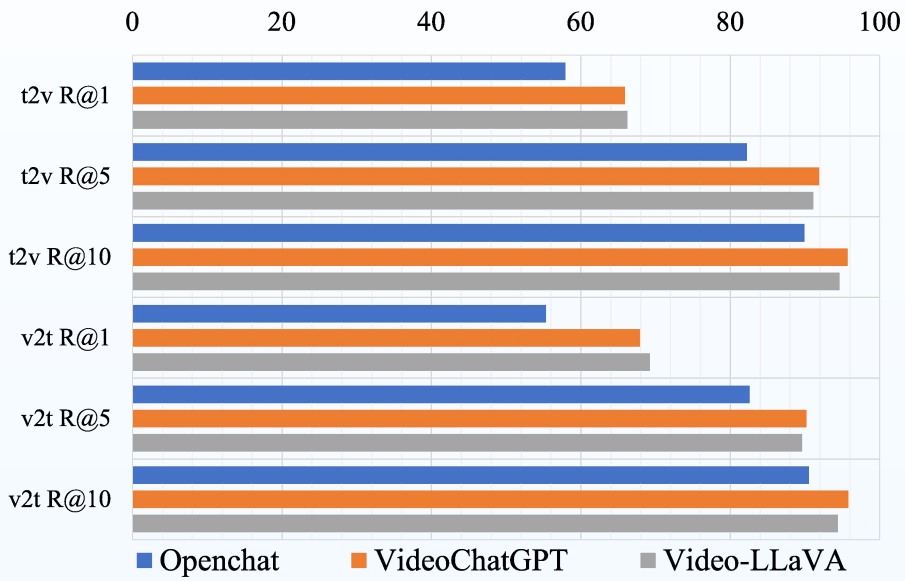
- ✓ Higher hit rate compared to the original
- ✓ Identifies more details like "**building**," "**dog**," and "**microphone**"

Supplement original texts with fine-grained details, enabling deeper video-text understanding

Experiments: Analytical Studies



	Prompt A	Prompt B	Prompt C
Complexity	Detailed	Medium	Simple
Supplement	✔	✘	✘
Detail	✔	✔	✘



- Superior results with **Prompt A & B**

⇨ ***Suggest lveraging mLLM as caption supplements rather than alternative expressions***

- Remain compressing the exisiting with **prompt C**

⇨ ***efficacy of leveraging mLLM in VTR***

- **Openchat (text-based LLM)** does not yield notable improvements.

⇨ ***Ineffectivenss of simple textual rewriting***

- Comparable results between **VideoChatGPT** and **Video-LLaVA**

⇨ ***Robustness to different mLLMs in the VTR task***

Conclusion & Future Work

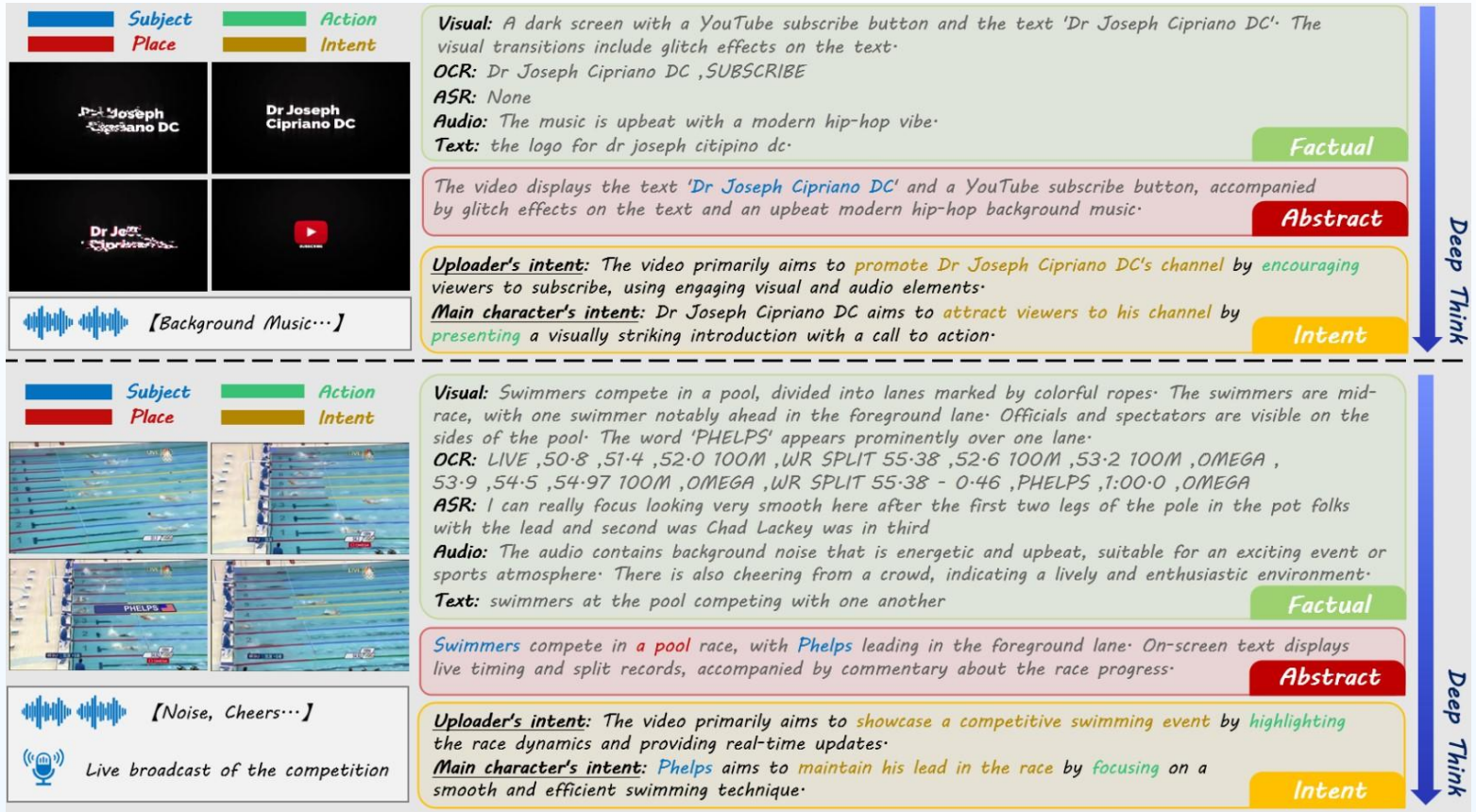
Key Contributions

- **mLLM as supplement:** Using mLLM to supplement video details rather than rewriting text to bridge cross-modal expression gaps
- **CEO method:** A Conical Embedding Optimization for video-text learning in compressed conical space
- **SOTA performance:** Achieved the best results on MSR-VTT, MSVD, and DiDeMo datasets
- **mLLM insights:** Provided analysis on mLLM selection and prompt design for VTR applications

Future Work

- **Auto generation:** Reduce reliance on the manual design of query prompts, automatically select the approciate prompts
- **Improved alignment:** Explore effective use of mLLM-supplemental captions for better cross-modal alignment
- **Larger dataset:** Enlarge the scale of dataset to facilitie video-text learning
- **Efficiency:** Reduce computation overhead of representation extraction and matching to achieve near-real-time retrieval

Advertise



Large number

110K video samples

Omni modality

Video+Audio+Text

Multiple layers

Factual-Abstract-Intent

VideoMind

An Omni-Modal Video Dataset with Intent Grounding

ECAI2025

Thank you for your attention

Baoyao Yang, Junxiang Chen and Wenbin Yao

廣東工業大學
Guangdong University of Technology

视频号

code available

Q&A *Welcome questions about our research*

{caryjxchen,wenbinyao}@tencent.com