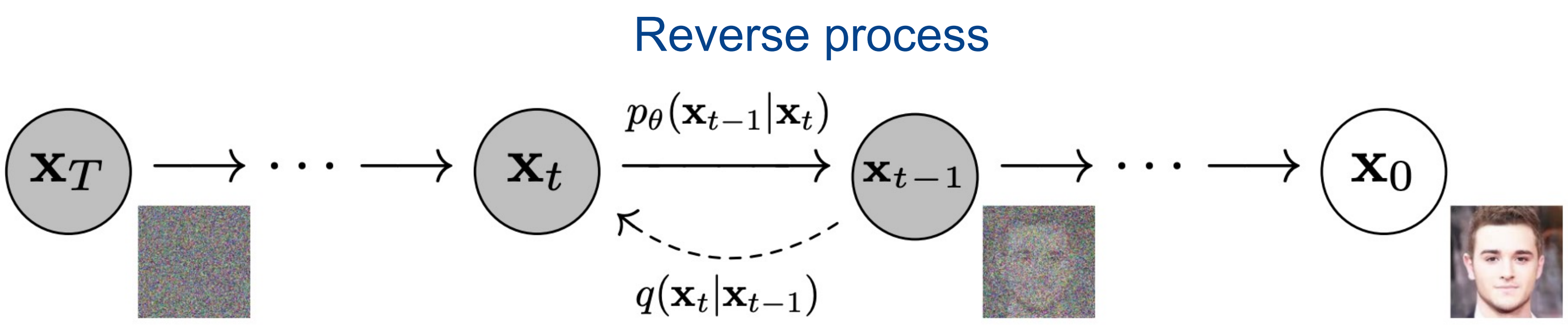
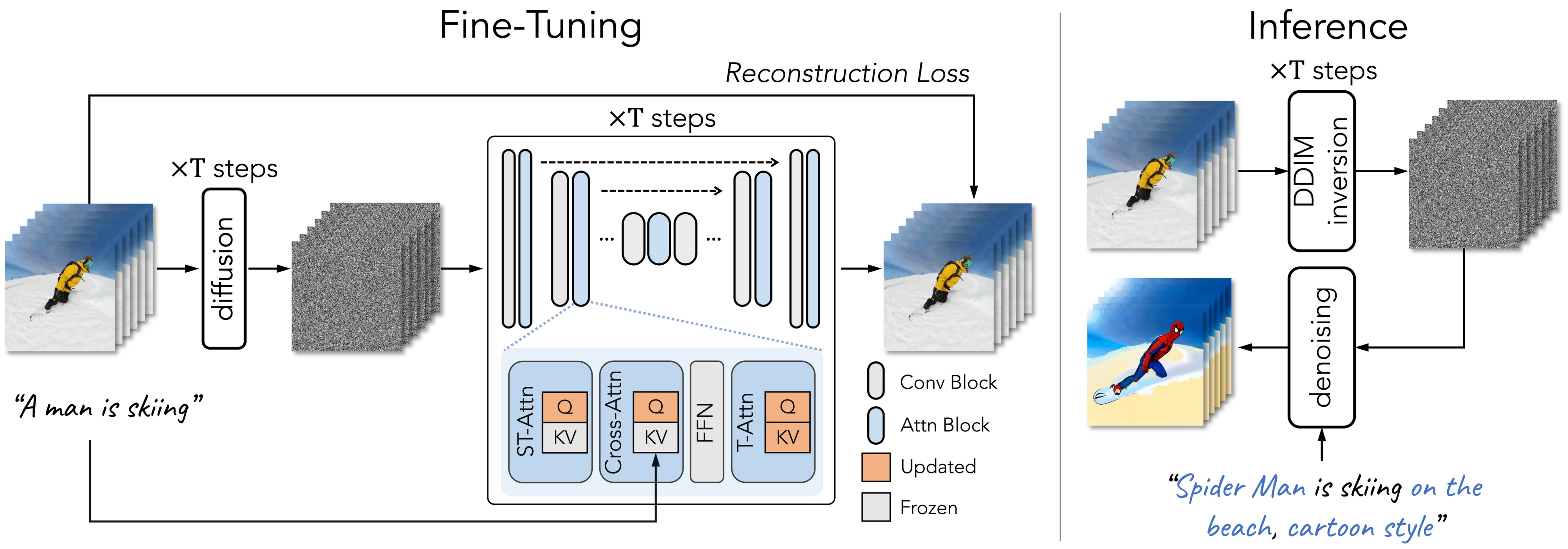
VideoGrain
Modulating Space-Time Attention for Multi-Grained Video Editing

Xiangpeng Yang[1], Linchao Zhu[2], Hehe Fan[2], Yi Yang[2]
[1]University of Technology Sydney, ReLER Lab; CCAI, [2]Zhejiang University

Preliminary – Diffusion Models





Preliminary – Video Editing

UTS UNIVERSITY OF TECHNOLOGY SYDNEY

- Development of foundational models

stability.ai sora CogVideo 腾讯混元-Video

- Various Video Editing:

Style editing

object editing

Motion Transfer

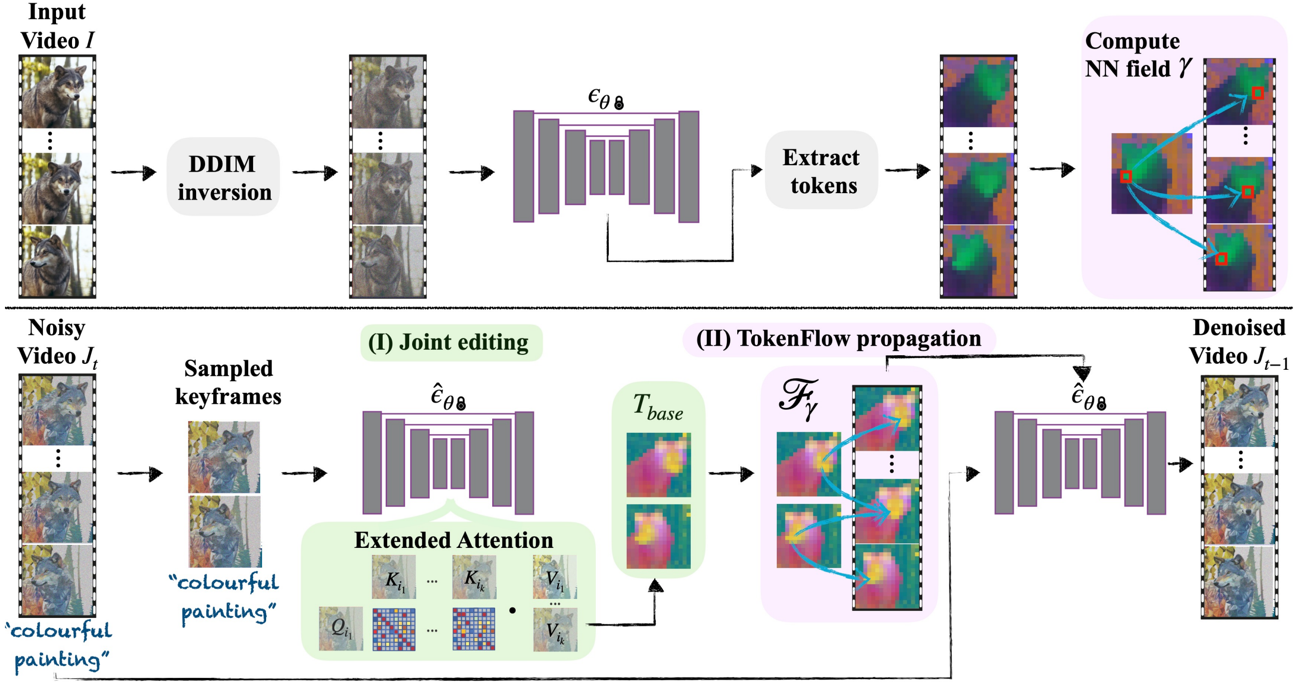
Motion Editing

Object insersion/inpainting

Video Editing

- Challenges
  - **时空一致性保持：**编辑需要考虑视频的时空连续性，包括光照条件（确保场景光影的自然过渡）、物体运动（维持对象的运动轨迹连贯性）、场景切换（实现流畅的镜头转场），以及全局一致性（保持整体视觉效果的统一）。如何在局部修改的同时维持全局的时空一致性？如何处理复杂场景下的动态物体编辑？
  - **语义理解与保持：**视频编辑需要深入的场景理解能力，涉及物体识别（准确定位和追踪目标对象）、语义分割（精确区分前景背景）、关系理解（把握物体间的交互关系），以及上下文连贯性（维持情节逻辑）。如何确保编辑操作符合场景的语义约束？如何在内容修改后保持原有的叙事逻辑？
  - **实时性与交互性：**专业的视频编辑要求良好的实时交互体验，包括即时预览（快速呈现编辑效果）、实时反馈（及时响应用户操作）。如何在有限的计算资源下实现流畅的编辑体验？如何平衡编辑精度和响应速度？

Video Editing-Style Editing

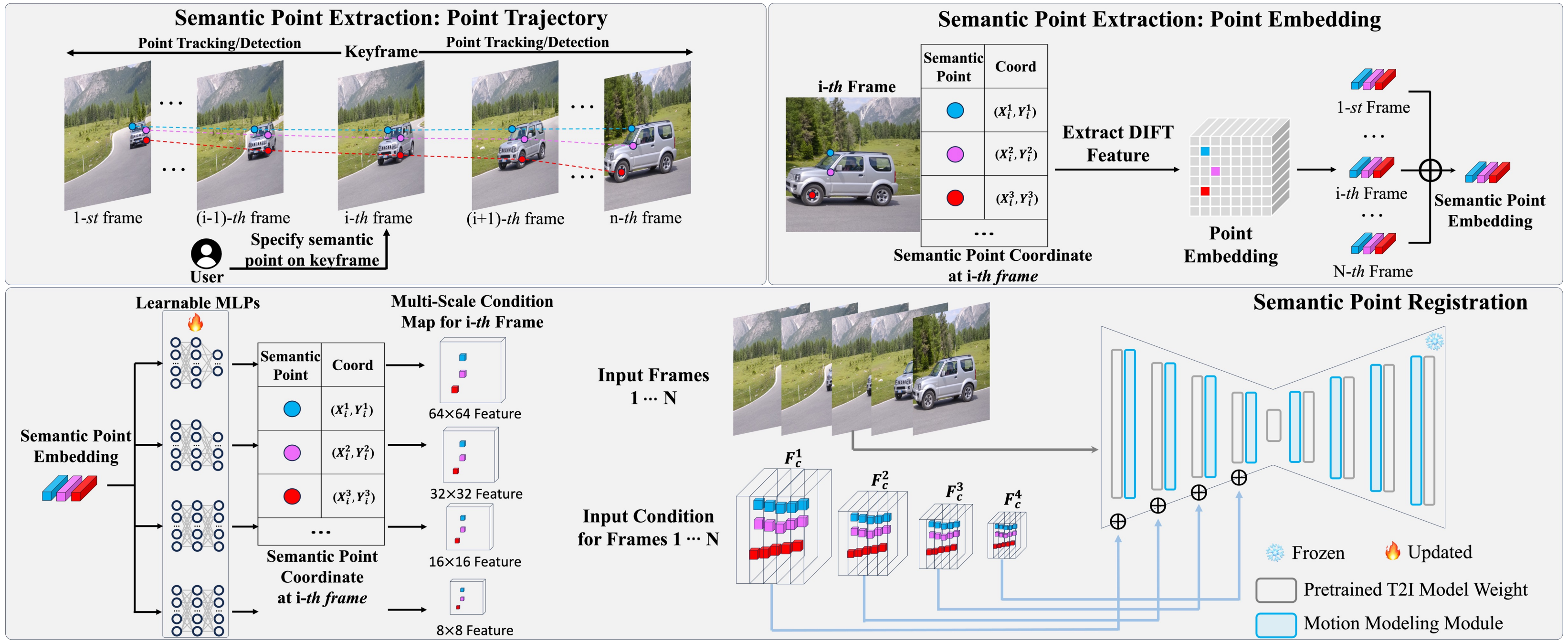


Token Flow, 基于文生图模型的视频编辑方案，通过编辑关键帧再利用最近邻区域回传特征的方法实现一致性的视频编辑

TokenFlow: Consistent Diffusion Features for Consistent Video Editing ICLR 2024

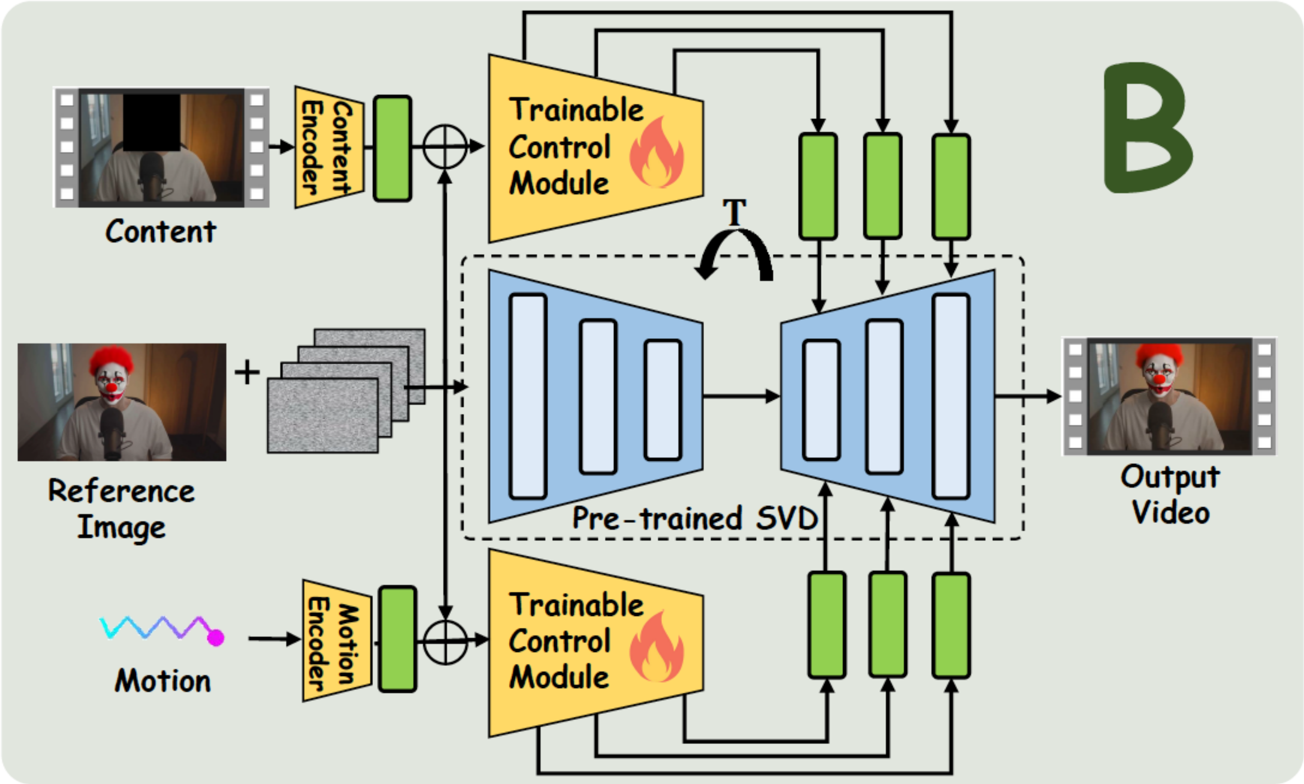
Video Editing-Shape Editing

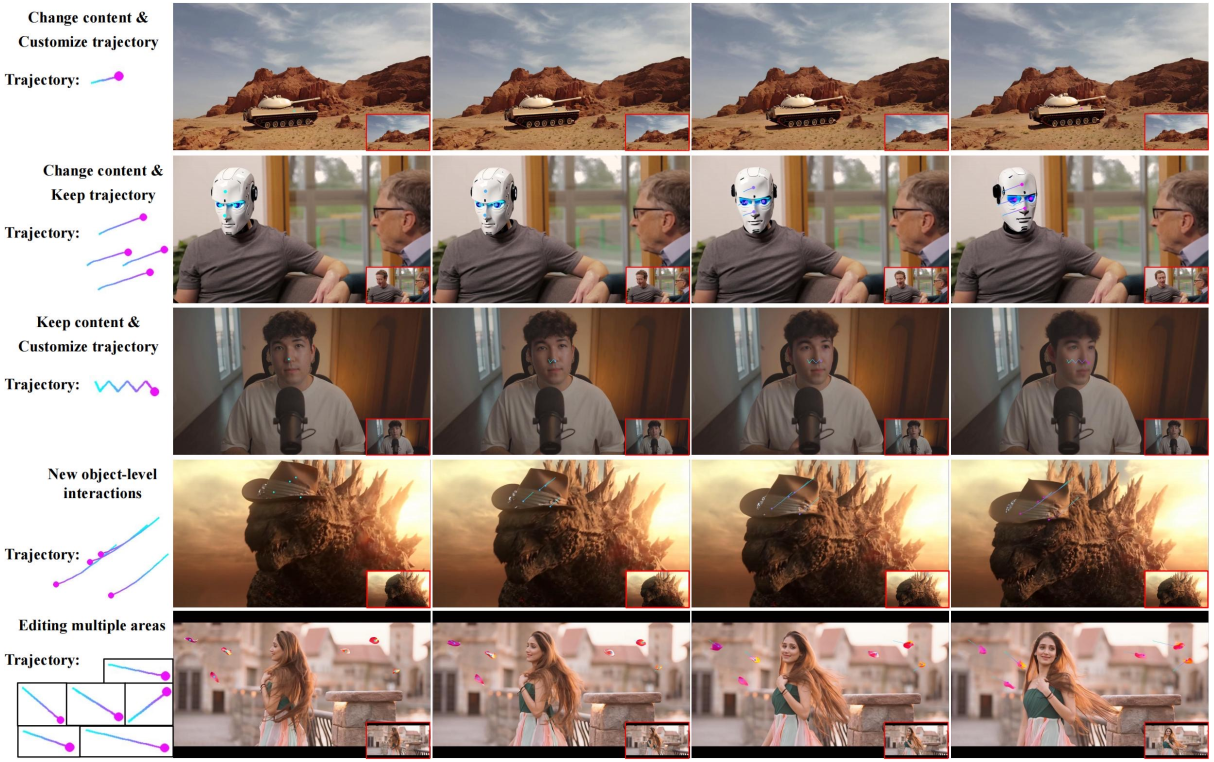
Video Swap, 基于扩散模型中semantic point的一致性， 通过用户可控的移除点或拖拽点的方式实现目标形状的编辑



VideoSwap: Customized Video Subject Swapping with Interactive Semantic Point Correspondence, CVPR 2024

Video Editing-Motion Editing





ReVideo, 基于图生视频的模型， 通过三阶段的预训练渐进式引入内容和动作的控制, 可以实现动作和内容的同时编辑

ReVideo: Remake a Video with Motion and Content Control, Neurips 2024

Video Editing-Motion Editing

Intro Video


ICLR
International Conference On
Learning Representations
VideoGrain
Modulating Space-Time Attention
for Multi-Grained Video Editing
ICLR 2025

VideoGrain – Task Definition

class level

instance level

part level: adding new objects

part level: modifying exisiting attributes

- **Class Level** : **Editing objects within the same class**
- **Instance Level**: **Editing each individual instance to distinct object**
- **Part Level**: **Applying part-level edit to specific elements of individual instances.**

grained level

Dense

Multi-Grained Video Editing

- **Objective: Multi-grained video Editing (Class/Instance/part level)**

Class Level | Instance Level | Part Level

VideoGrain

source video | class level | instance level | part level

Previous SOTA Performance

TokenFlow | GroundVideo | DMT | Pika

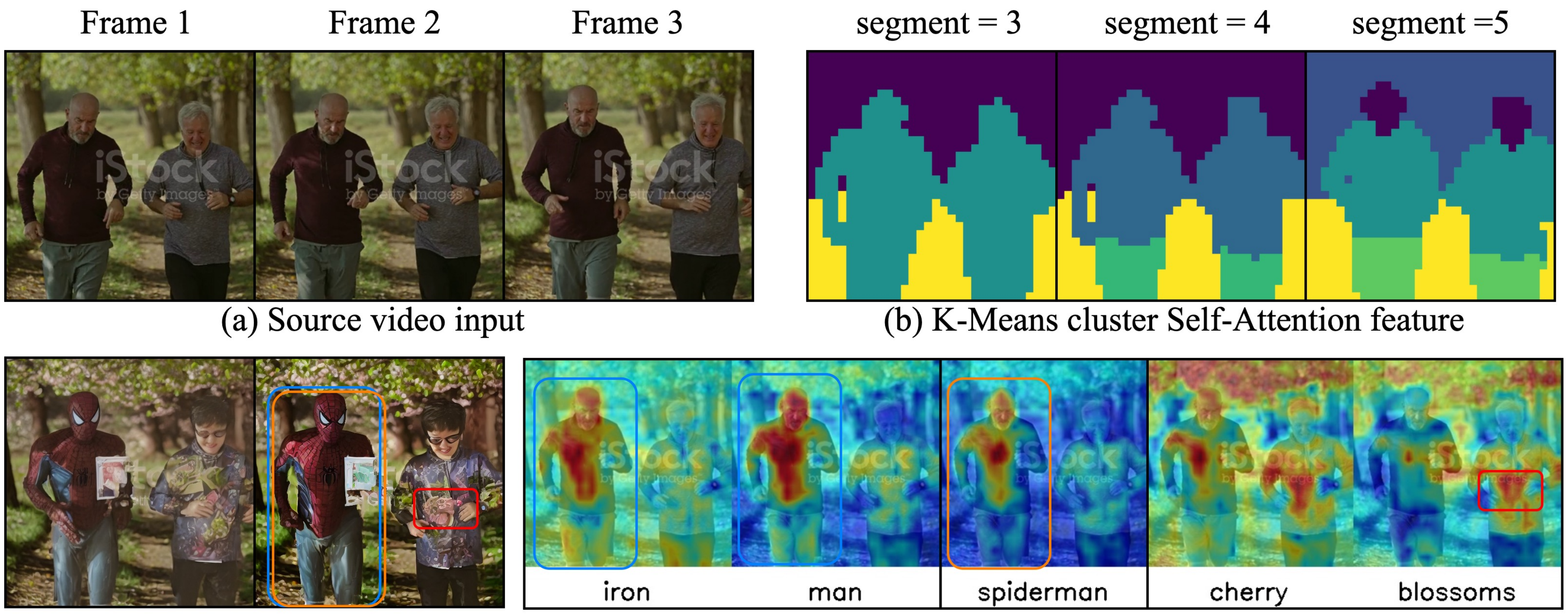
VideoGrain: Modulating Space-Time Attention for Multi-Grained Video Editing, ICLR 2025

Challenges and Motivation

Challenges: Direct editing leads to failed edit and attention leakage.

Reasons:

- **Feature coupling:** Diffusion models cannot distinguish between left and right instances. Increasing clusters only refines the layout but fails to **separate instances**.
- **Text-to-region control:** Editing fails due to **inaccurate cross-attention weights**. Proper editing should ensure accurate attention distribution across regions.



(a) Source video input (b) K-Means cluster Self-Attention feature

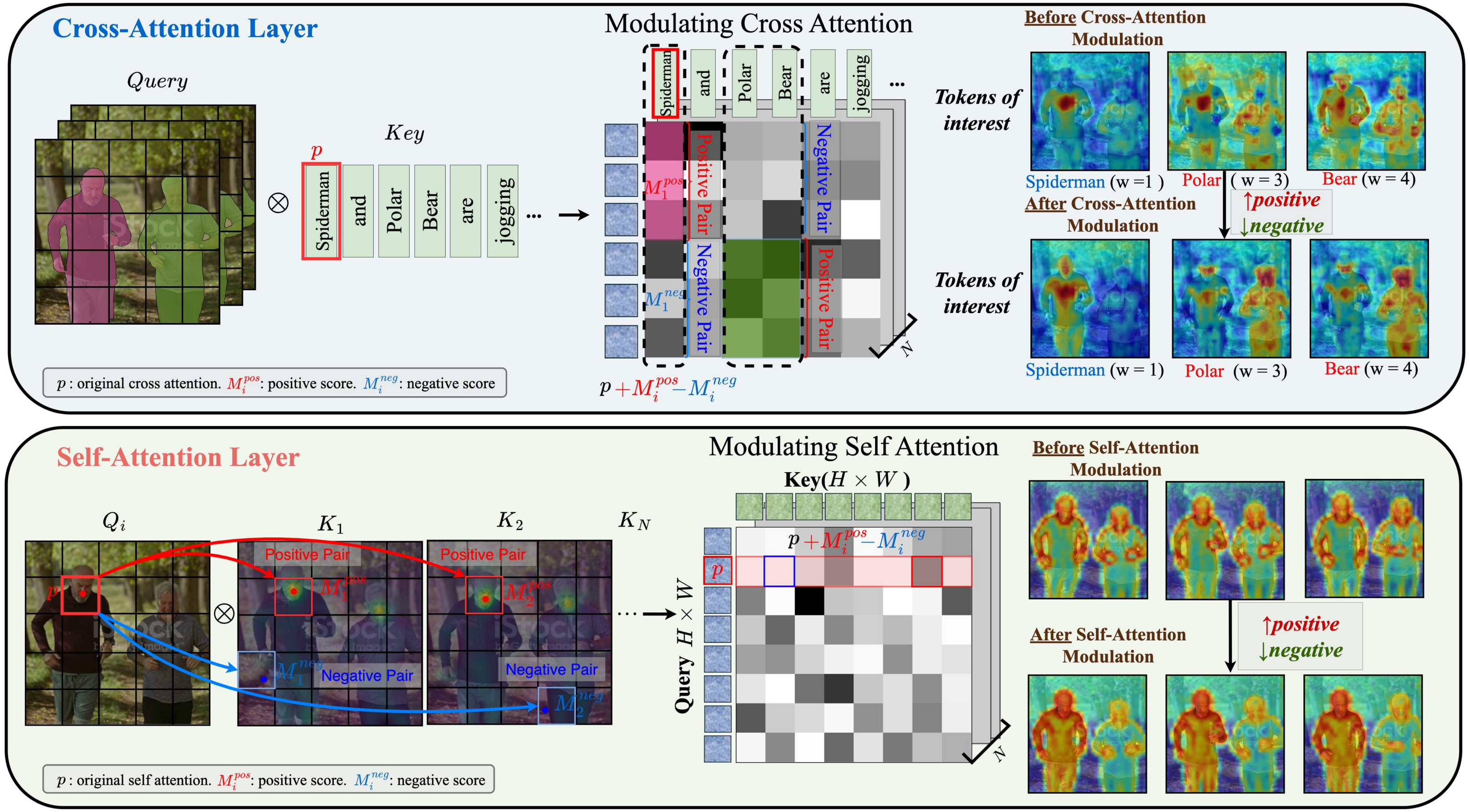
(c) Instance-level failed case (d) Cross-Attention Map: "*An Iron Man and a Spiderman are jogging under cherry blossoms*"

Framework

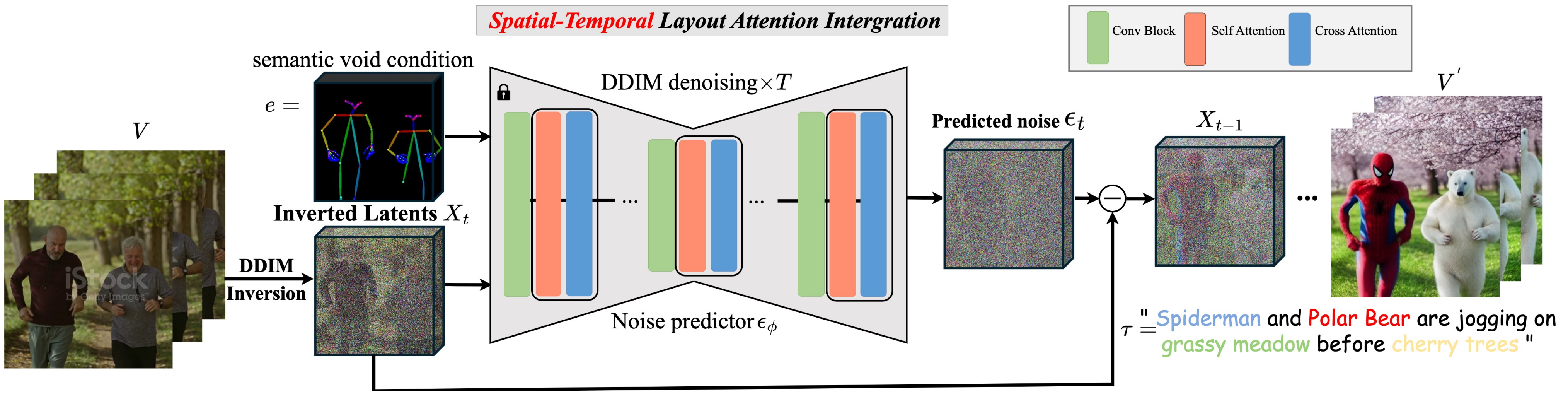
UNIVERSITY OF TECHNOLOGY SYDNEY

Unified ⬆️increase positive, ⬇️decrease negative manner:

- **Text-to-region control**: Each local prompt and its location as positive pairs, while the prompt and outside-location areas are negative pairs,
- **Keep feature Separation:** Enhance positive awareness within intra-regions, restrict negative interactions between inter-regions across frames.



Framework



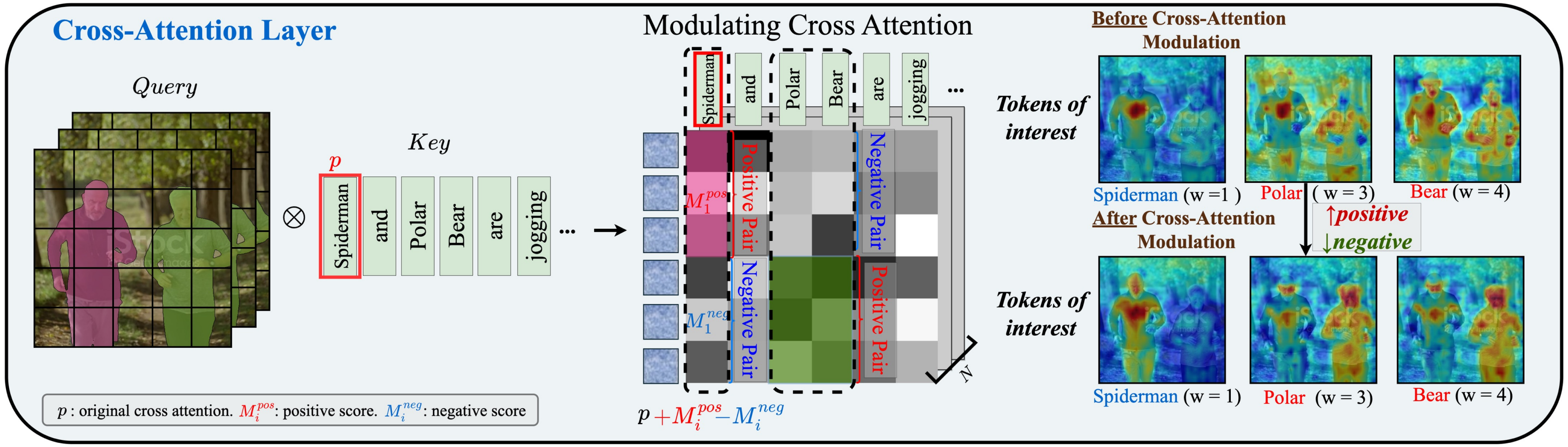
- Modulated Attention

$$A_i^{\text{self/cross}} = \text{softmax}\left(\frac{QK^\top + \lambda M^{\text{self/cross}}}{\sqrt{d}}\right),$$

- Query-key condition map

$$M^{\text{self/cross}} = R_i \odot M_i^{\text{pos}} - (1 - R_i) \odot M_i^{\text{neg}},$$

Modulate Cross-Attention for Text-to-Region Control



- Modulated Cross Attention

$$A_i^{\text{self/cross}} = \text{softmax}\left(\frac{QK^\top + \lambda M^{\text{self/cross}}}{\sqrt{d}}\right),$$

- Cross-attn qk condition map

$$M^{\text{self/cross}} = R_i \odot M_i^{\text{pos}} - (1 - R_i) \odot M_i^{\text{neg}},$$

- Positive/Negative value definition

$$M_i^{\text{pos}} = \max(QK^\top) - QK^\top,$$
$$M_i^{\text{neg}} = QK^\top - \min(QK^\top),$$

- Regularize cross condition map

$$R_i^{\text{cross}}[x, y] = \begin{cases} m_{i,k}, & \text{if } y \in \tau_k \\ 0, & \text{otherwise} \end{cases},$$

Modulate Self-Attention to Keep Feature Separation

- Modulated self-attention

$$A_i^{\mathrm{self/cross}} = \mathrm{softmax}\left(\frac{QK^\top + \lambda M^{\mathrm{self/cross}}}{\sqrt{d}}\right),$$

- Self-attn qk condition map

$$M^{\mathrm{self/cross}} = R_i \odot M_i^{\mathrm{pos}} - (1 - R_i) \odot M_i^{\mathrm{neg}},$$

- Positive/Negative value definition

$$M_i^{\mathrm{pos}} = \max(Q_i[K_1, \cdots, K_n]^\top) - Q_i[K_1, \cdots, K_n]^\top),$$
$$M_i^{\mathrm{neg}} = Q_i[K_1, \cdots, K_n]^\top - \min(Q_i[K_1, \cdots, K_n]^\top).$$

- Regularize self condition map

$$R_i^{\mathrm{self}}[x, y] = \begin{cases} 0, \forall j \in [1 : N], \text{if } m_{i,k}[x] \neq m_{j,k}[y] \\ 1, \text{otherwise} \end{cases}.$$

Qualitative Results

UTS UNIVERSITY OF TECHNOLOGY SYDNEY

- **Solely edit** -> joint edit, background unchanged.
- **Instance level:** human/animal instances, complex motion and multi-region editing
- **Part Level**: adding new object, modify part-level attribute.

solely edit -> joint edit　　animal instances

complex motion　　multi-region editing　　riding bikes

part-level adding sunglasses　　part-level color change

VideoGrain – Comparison with SOTA

UTS UNIVERSITY OF TECHNOLOGY SYDNEY

part level: 索尔带着墨镜在夜晚挥舞着红色的拳击手套

human instances: 钢铁侠和一个猴子在雪地上樱花树下骑车

animal instances: 一个熊猫和一个贵宾犬在星月夜的草地上玩耍

Quantitative result

UNIVERSITY OF TECHNOLOGY SYDNEY

Quantitative comparison

Method	Automatic Metric				Human Evaluation		
	CLIP-F ↑	CLIP-T ↑	Warp-Err ↓	Q-edit ↑	Edit-Acc ↑	Temp-Con ↑	Overall ↑
FateZero	95.75	33.78	3.08	10.96	59.8	78.6	59.6
ControlVideo	97.71	34.41	4.73	7.27	53.2	50.0	43.6
TokenFlow	96.48	34.59	2.82	12.28	45.4	50.4	39.8
Ground-A-Video	95.17	35.09	4.43	7.92	69.0	72.0	63.2
DMT	96.34	34.09	2.05	16.63	58.7	79.4	64.5
VideoGrain(ours)	**98.63**	**36.56**	**1.42**	**25.75**	**88.4**	**85.0**	**83.0**

Table 1: Quantitative comparison of automatic metrics and human evaluation. The best results are **bolded**.

	Time(min) ↓	Memory (GB) ↓	RAM (GB) ↓
FateZero	8.68	27.35	144.22
ControlVideo	4.41	16.15	7.03
TokenFlow	4.56	17.84	5.35
Ground-A-Video	5.81	17.31	9.96
DMT	5.79	27.88	8.12
VideoGrain	**3.83**	**15.94**	**4.42**

Table 2: Efficiency comparison.

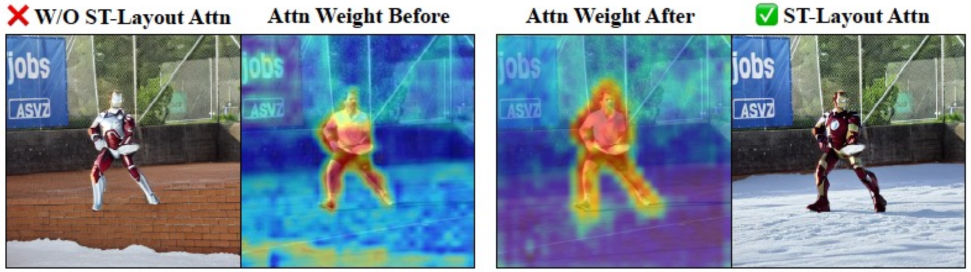


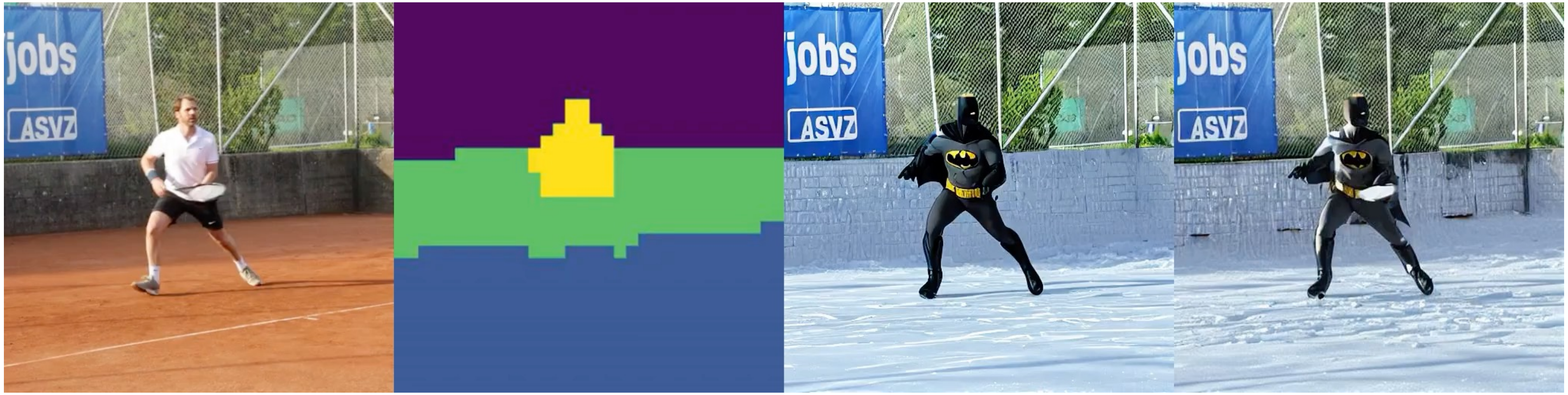
Figure 7: Attention weight distribution.

Ablation study

UTS UNIVERSITY OF TECHNOLOGY SYDNEY

source video perf-rame fist+last frame VideoGrain(full frame)

Temporal focus



Wo-SAM-Track Masks

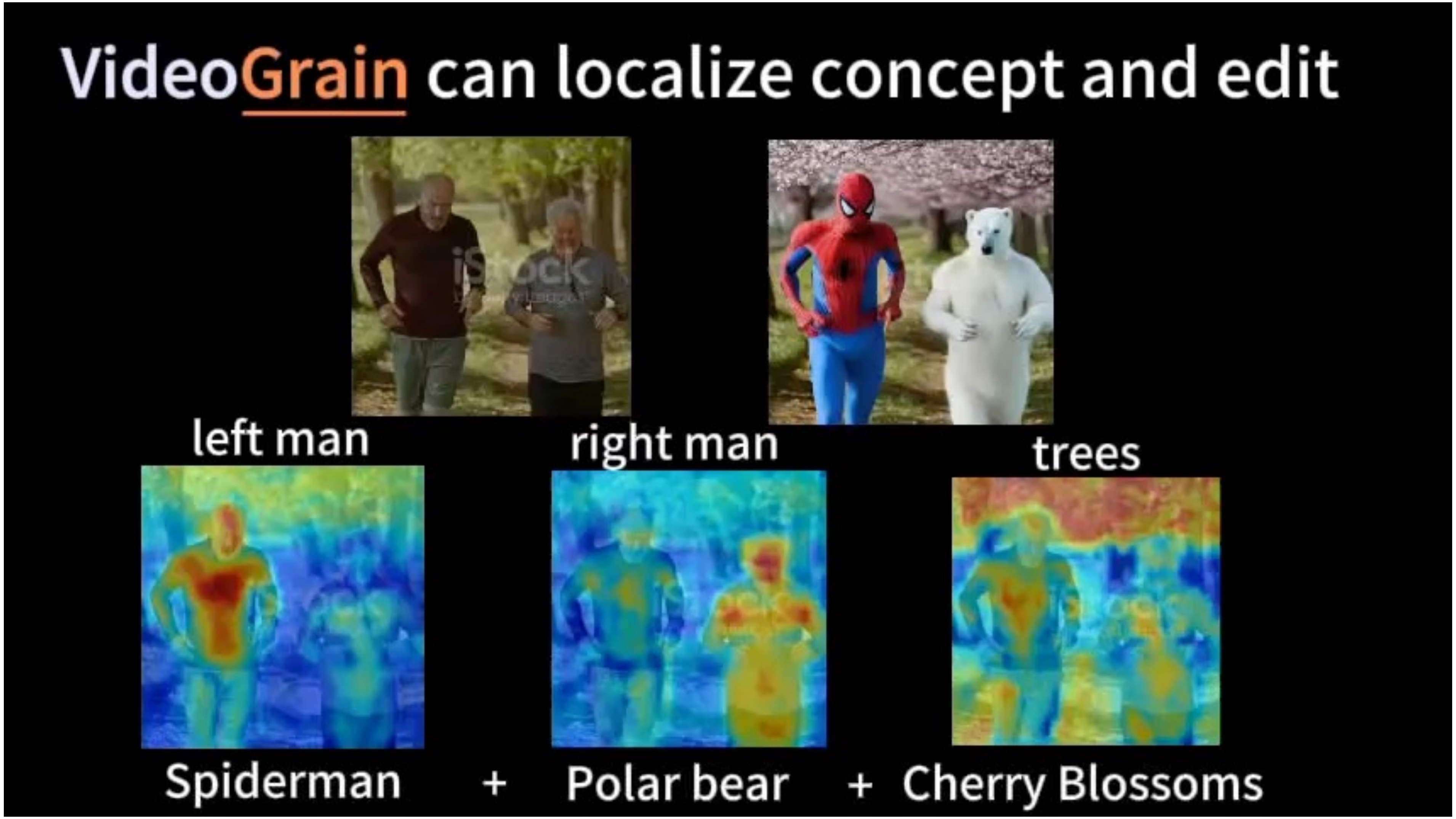
Ablation study – Other method + mask

Source prompt: red man and gray man are jogging under green trees
Edit prompt: Spider Man and Polar Bear are jogging under cherry blossoms
VideoP2P setting: Attention Replace 3 subject words + Attention Reweight
(Spider man: 4, polar bear: 4,cherry blossoms:2)

Localize concept and edit

Localize concept and edit:

- VideoGrain 可以定位“概念”，并在一次去燥中实现多个概念的同时编辑



VideoGrain: Modulating Space-Time Attention for Multi-Grained Video Editing, ICLR 2025

Future direction

- Video Editing
  - Motion Editing
  - Efficiency
  - Unified Understanding and Editing (4o image gen -> understanding intend video editing)
- Video Generation
  - Multi-view Video Generation
  - Movie-level multi-shot video gen

UTS UNIVERSITY OF TECHNOLOGY SYDNEY

Thanks

XIANGPENG YANG

ReLER