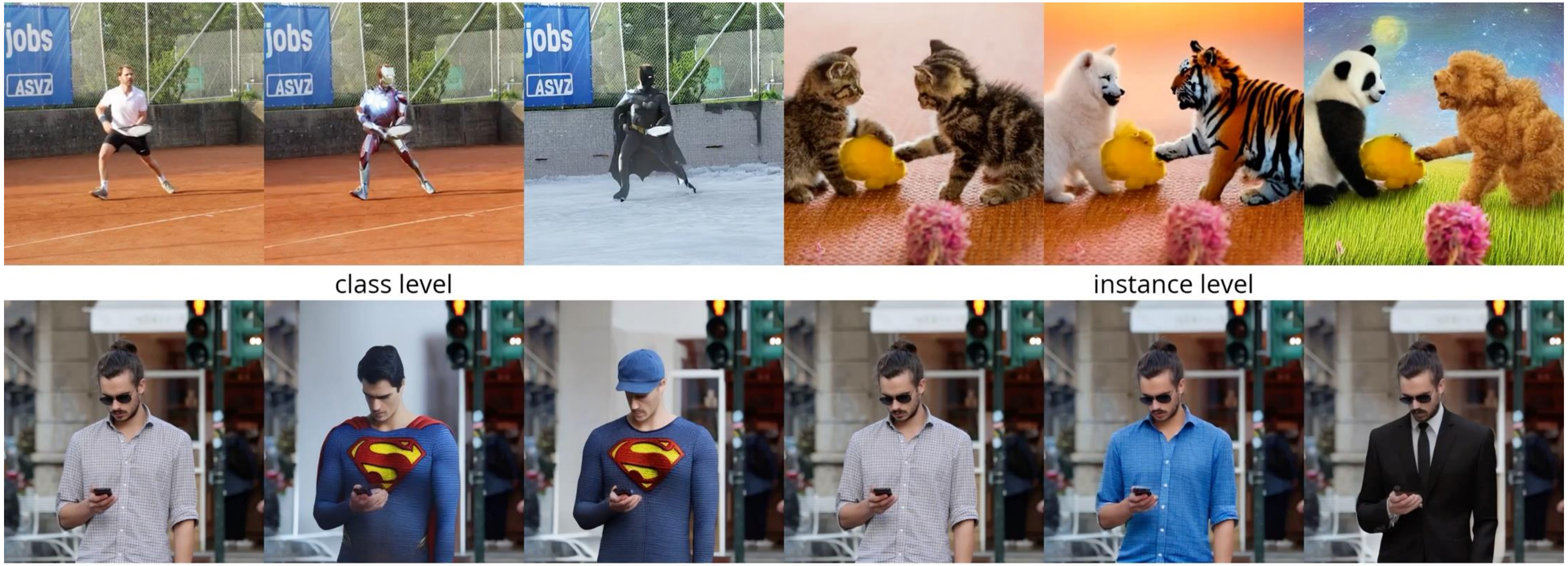
VideoGrain
Modulating Space-Time Attention for Multi-Grained Video Editing

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

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

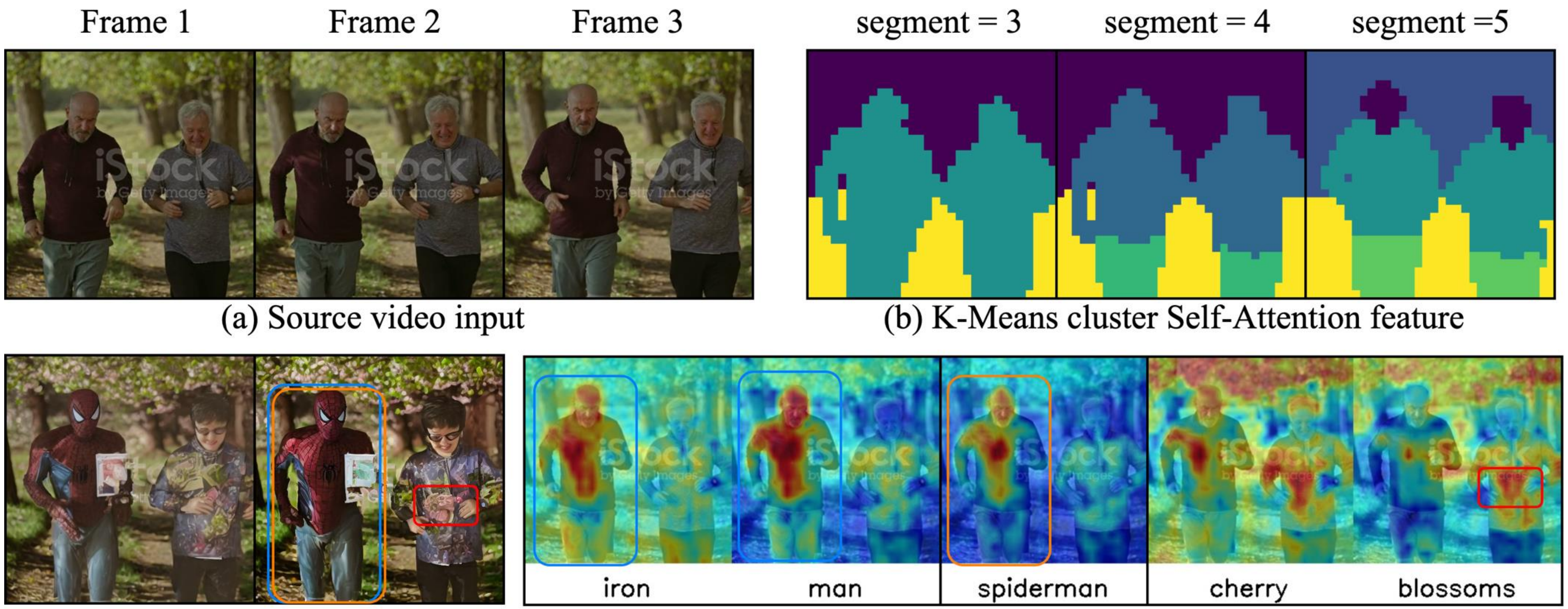
VideoGrain: Modulating Space-Time Attention for Multi-Grained Video Editing, ICLR 2025

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.



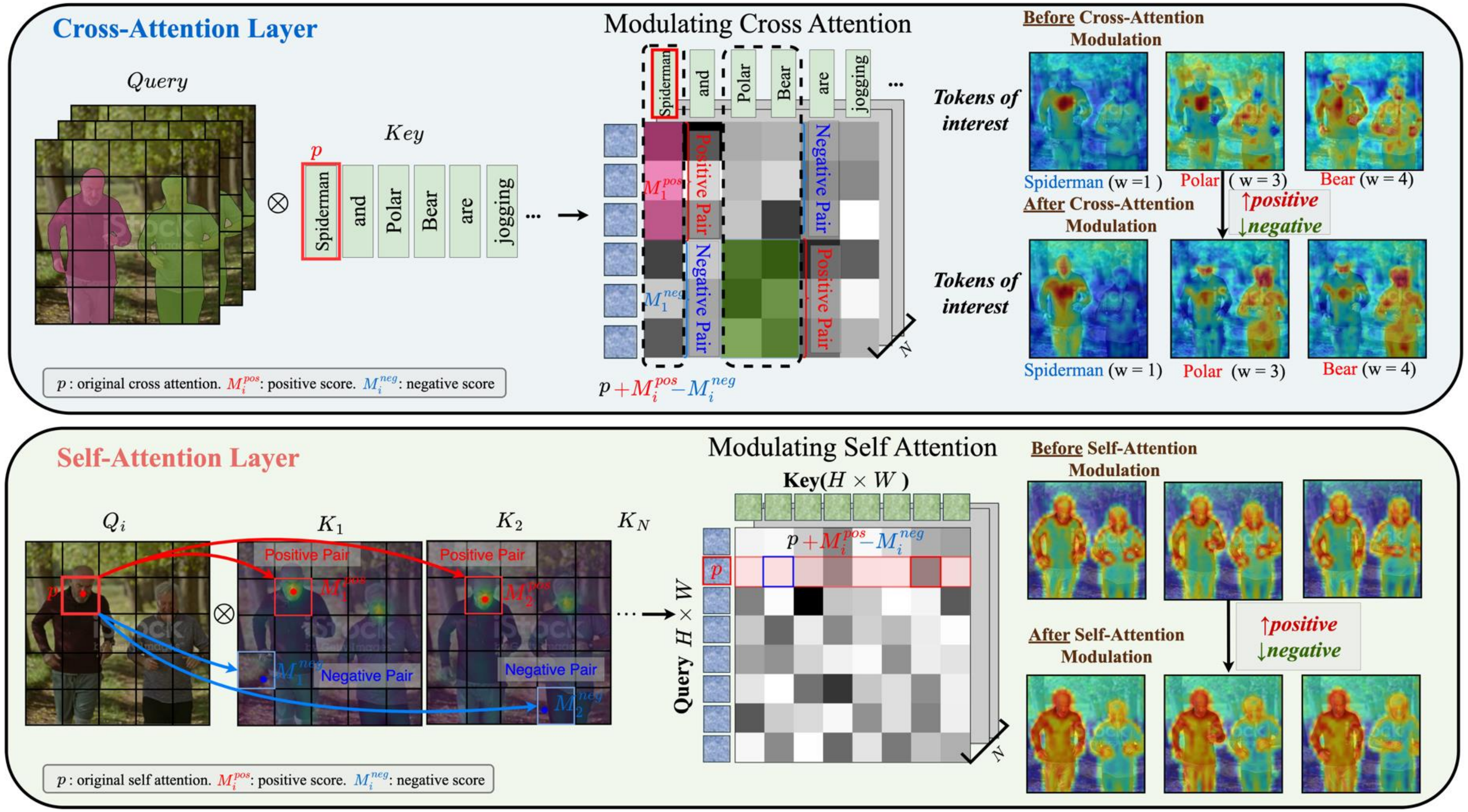
(a) Source video input (b) K-Means cluster Self-Attention feature

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

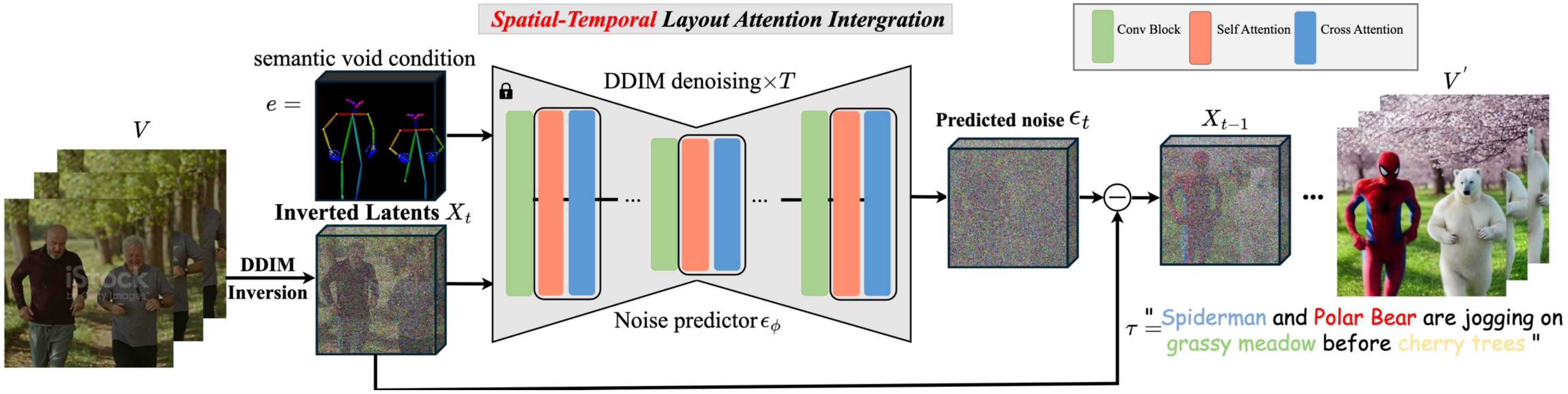
Framework

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



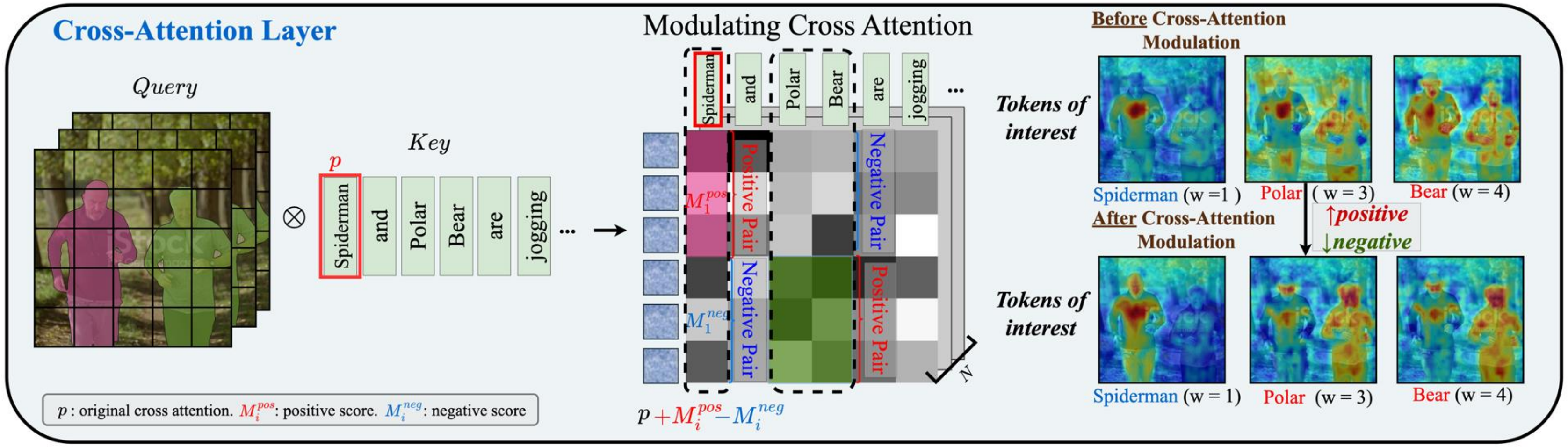
- Modulated Attention

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

- 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

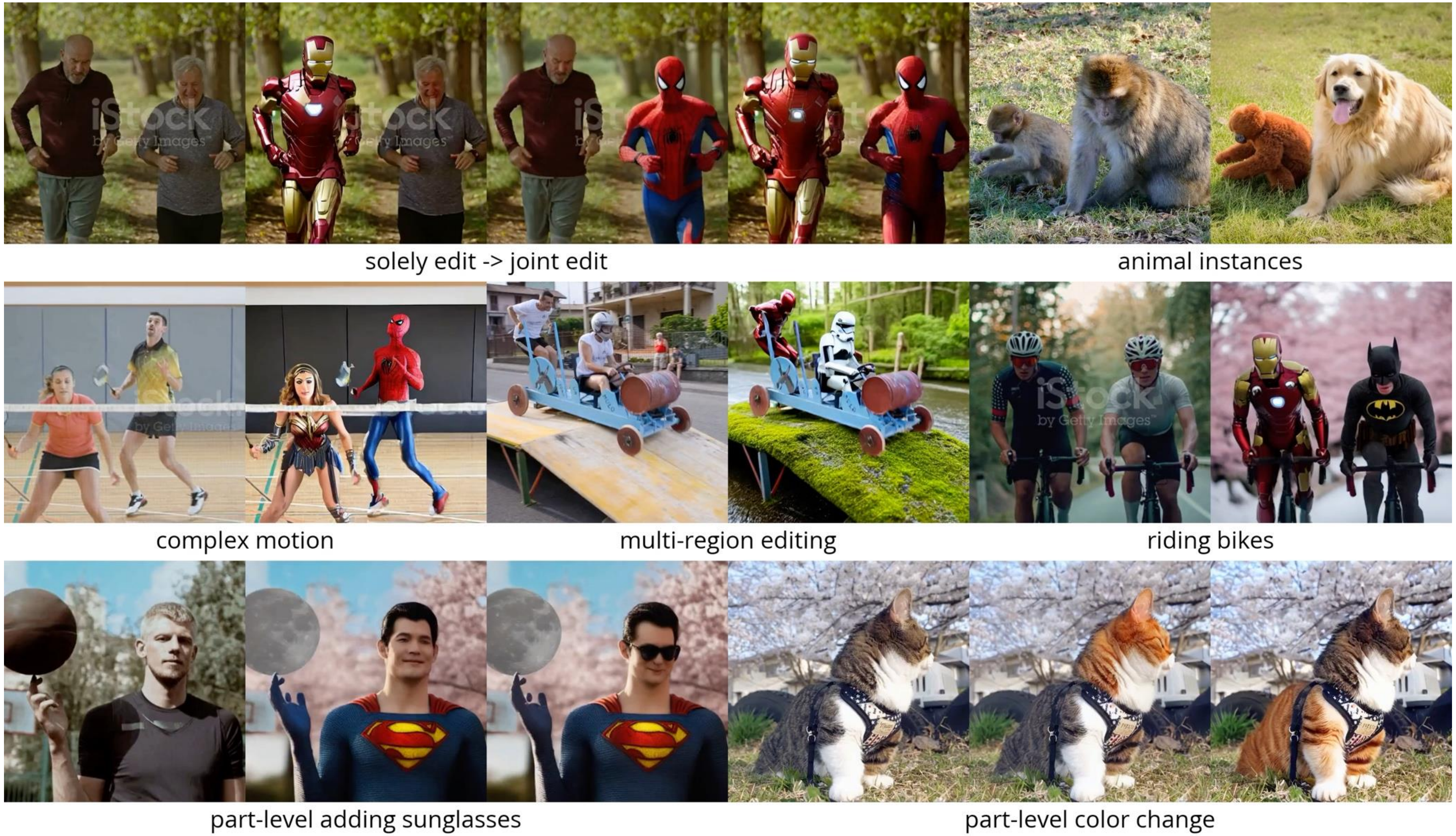
$$M_i^{\mathrm{pos}} = \max(Q_i[K_1, \cdots, K_n]^\top) - Q_i[K_1, \cdots, K_n]^\top),$$
$$M_i^{\mathrm{neg}} = Q_i[K_1, \cdots, K_n]^\top - \min(Q_i[K_1, \cdots, K_n]^\top).$$

- 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

- **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

Quantitative result

Quantitative comparison

Method	Automatic Metric CLIP-F ↑	CLIP-T ↑	Warp-Err ↓	Q-edit ↑	Human Evaluation 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.

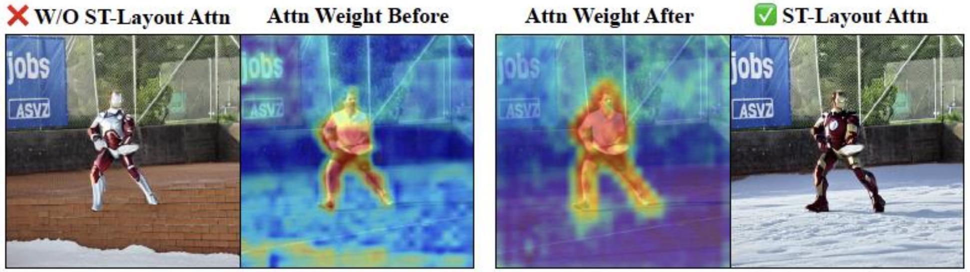


Figure 7: Attention weight distribution.

Ablation study

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

Temporal focus

Wo-SAM-Track Masks

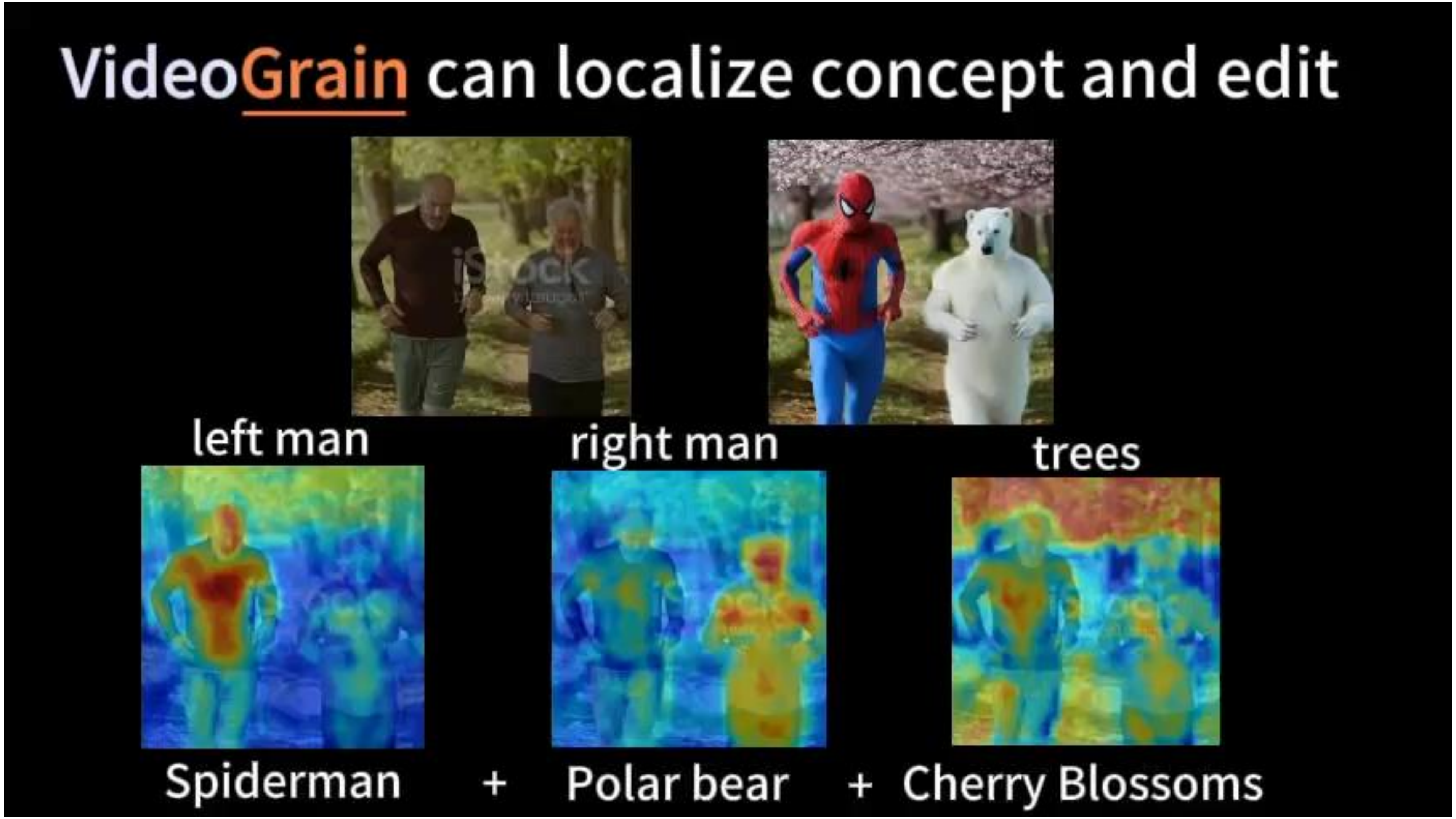
Ablation study – Other method + mask

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 can localize concept and edit multi-concept in one denoising process.



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

Thanks

XIANGPENG YANG

ReLER