The Hong Kong Polytechnic University
香港理工大學

Poster ID: WED-AM-219

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

# MDQE: Mining Discriminative Query Embeddings to Segment Occluded Instances on Challenging Videos

Minghan LI, Shuai LI, Wangmeng Xiang, and Lei Zhang*

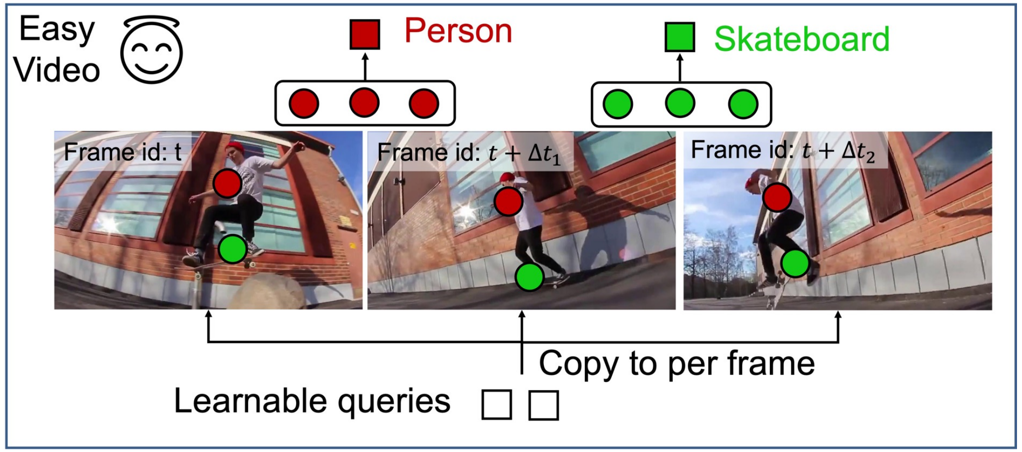
# 1. Motivations

➢ **Architecture of previous clip-based VIS methods**

To distinguish objects depends mainly on:

Positions + Categories

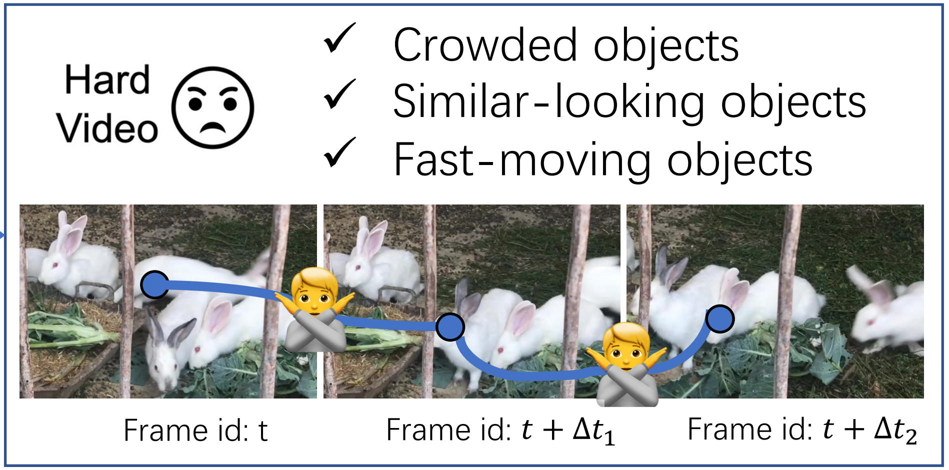
**SeqFormer:** (ECCV 2022)




➢ **Our work: mining discriminative object embeddings**

- Embedding initialization for object tokens
- Inter-instance mask repulsion loss

poor temporal consistency




Which is the target rabbit in the following frames?

# 2. Methodology

**To Mine discriminative object embeddings:**

2.1 Query initialization for object tokens

2.2 Inter-instance mask repulsion loss

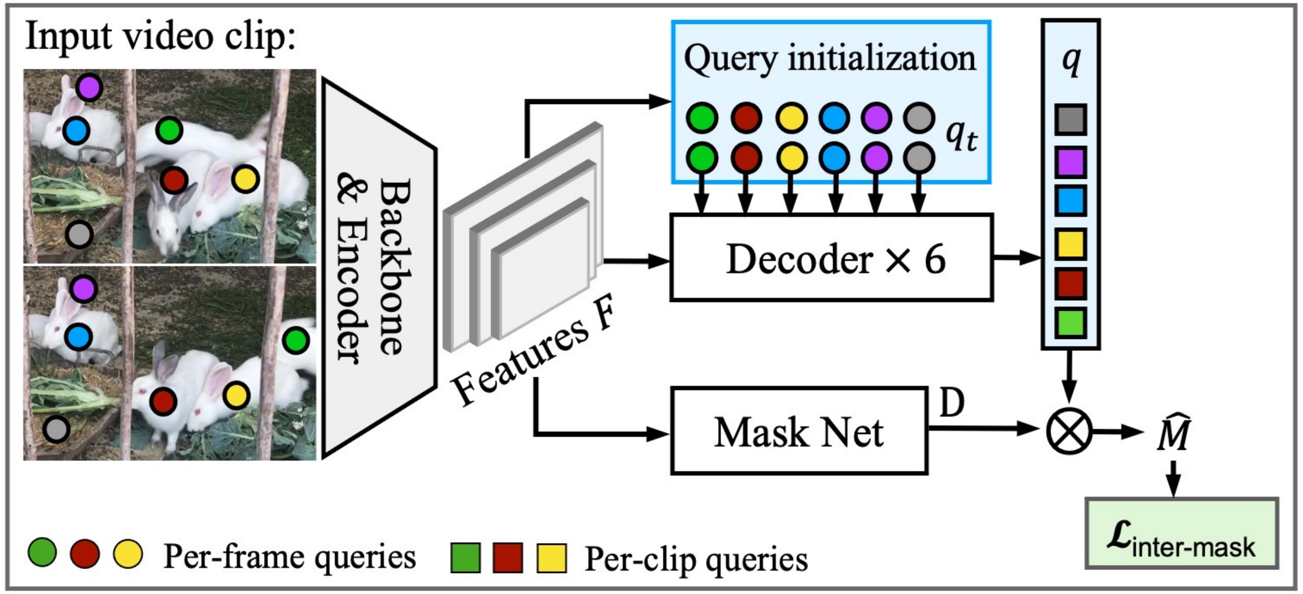
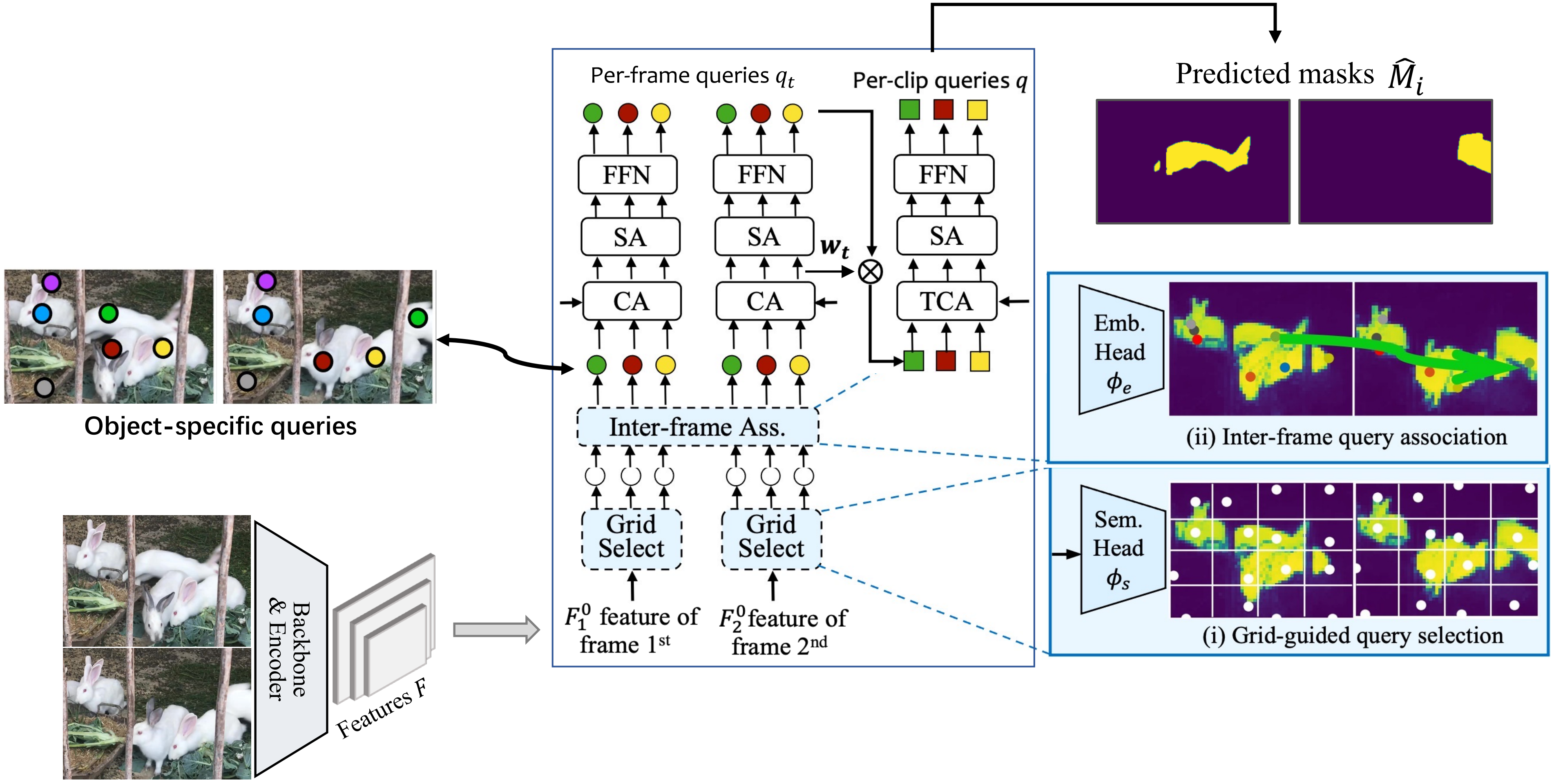



Fig. 1 Overview architecture of our proposed MDQE

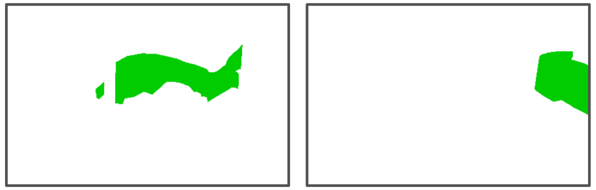
# 2.1 Query Initialization

Architecture of the first decoder layer with our proposed query token initialization:



# 2.2 Inter-instance Mask Repulsion Loss

I. Ground-truth masks $M_i$



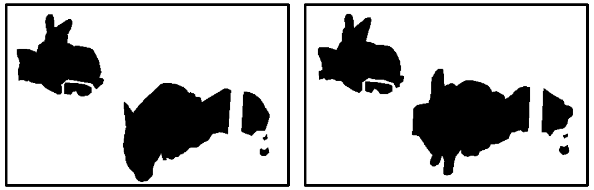
Predicted mask $\widehat{M}_i$

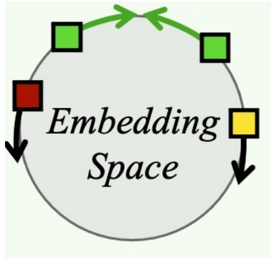
II. Define its nearby non-target instances via box IoU:

$$o_i = \{j \mid \max_{t \in [1,T]} \mathrm{IoU}(B_{ti}, B_{tj}) > \epsilon, \forall j \in [1, K], j \neq i\},$$

III. Complementary GT inter-instance mask:

$$M_{o_i} = \cup_{j \in o_i} M_j,$$






The formula of the inter-instance BCE loss is:

$$\mathcal{L}_{\textbf{BCE-inter}} = \frac{1}{|W_i|} \sum_{p=1}^{N} \boxed{W_{ip}} \mathrm{BCE}(\hat{M}_{ip}, M_{ip}), \quad (4)$$

where $p$ is the pixel position index. if $M_{ip} \cup M_{o_i p} = 1$, $W_{ip}$ = 2 otherwise 1.

The formula of inter-instance Dice loss is:

$$\mathcal{L}_{\textbf{Dice-inter}} = 1 - \frac{2|\hat{M}_i \odot M_i| + \boxed{|(1 - \hat{M}_i) \odot M_{o_i}|}}{|\hat{M}_i| + |M_i| + \boxed{|M_{o_i}|}} \quad (5)$$

# 3. Experimental Results

## 3.1 Ablation study

## 3.2 Main results

## 3.3 Visualization

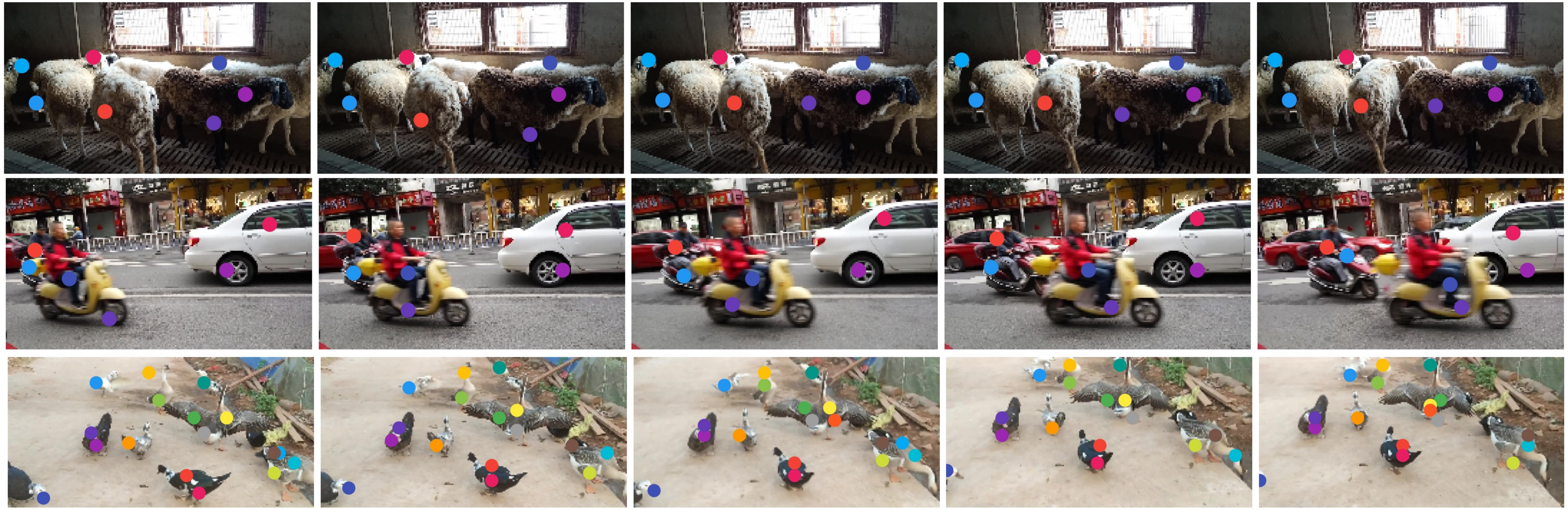
# 3.1 Ablation study for query initialization

| | Init. | Arch. | TCA | mAP | $AP_{50}$ | $AP_{75}$ | $AP_{so}$ | $AP_{mo}$ | $AP_{ho}$ |
|---|---|---|---|---|---|---|---|---|---|
| | | I2O | | 15.4 | 31.3 | 14.3 | 31.8 | 17.3 | 3.2 |
| | ✓ | I2O | | 19.8 | 40.6 | 18.2 | 36.3 | 22.6 | 6.5 |
| | ✓ | O2I | | 24.2 | 47.5 | 22.9 | 40.9 | 27.3 | 8.4 |
| +9.8% | ✓ | O2I | ✓ | 25.6 | 49.1 | 24.9 | 41.9 | 29.0 | 11.2 |

(a) Initialization for frame-level queries.

| $w$ | Assoc. | mAP | $AP_{50}$ | $AP_{75}$ | $AP_{so}$ | $AP_{mo}$ | $AP_{ho}$ |
|---|---|---|---|---|---|---|---|
| 0 | | 28.5 | 53.0 | 26.9 | 47.6 | 32.5 | 11.9 |
| 3 | ✓ | 29.7 | 55.6 | 27.1 | 48.9 | 34.5 | 12.2 |
| 5 | ✓ | 30.6 | 57.2 | 28.2 | 49.3 | 35.1 | 13.6 |
| 7 | ✓ | 30.5 | 57.1 | 28.6 | 49.1 | 33.7 | 13.7 |

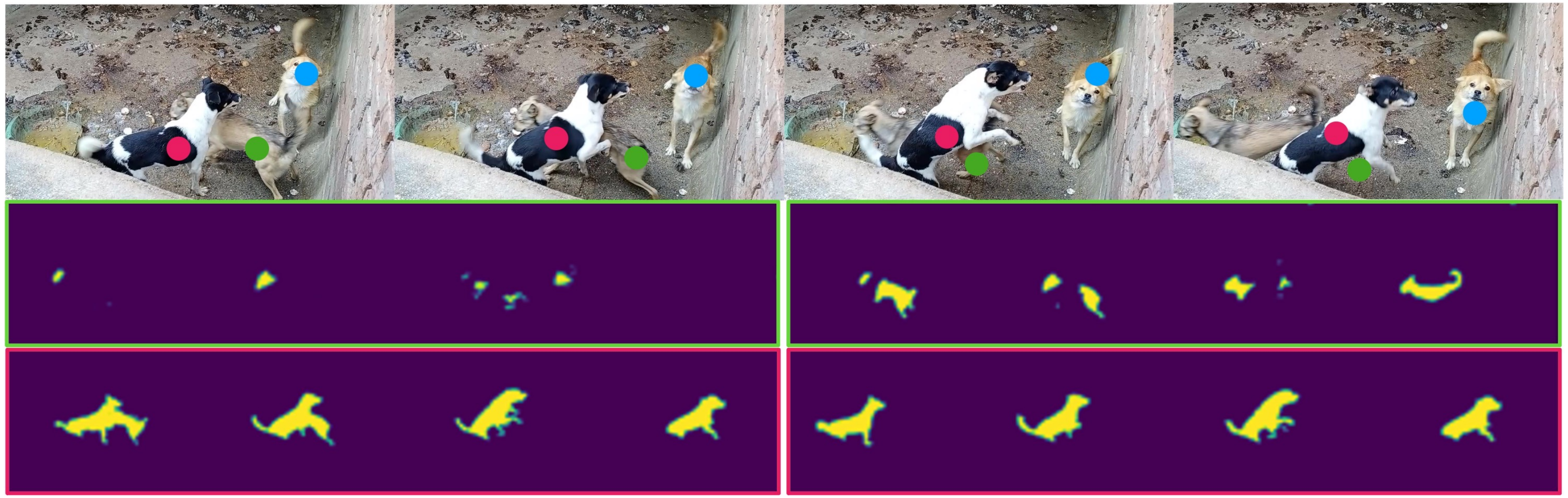
(b) Inter-frame query association, where $w$ controls the window size.



# 3.1 Ablation study for mask repulsion loss

| $\mathcal{L}_{\text{BCE-inter}}$ | $\mathcal{L}_{\text{Dice-inter}}$ | $\epsilon$ | mAP | $AP_{50}$ | $AP_{75}$ | $AP_{so}$ | $AP_{mo}$ | $AP_{ho}$ |
|---|---|---|---|---|---|---|---|---|
| | | | 29.0 | 51.6 | 29.5 | 44.7 | 31.3 | 11.8 |
| 2 | | 0.1 | 30.5 | 55.6 | 29.5 | 46.7 | 33.1 | 12.9 |
| 2 | ✓ | 0.1 | 31.2 | 56.8 | 30.4 | 48.6 | 34.5 | 13.5 |
| 2 | ✓ | 0.5 | 30.9 | 56.4 | 30.5 | 47.2 | 34.2 | 13.3 |

(c) Inter-instance mask repulsion loss.



(a) Typical mask prediction loss (b) Our inter-instance mask repulsion loss

# 3.1 Ablation study for clip length

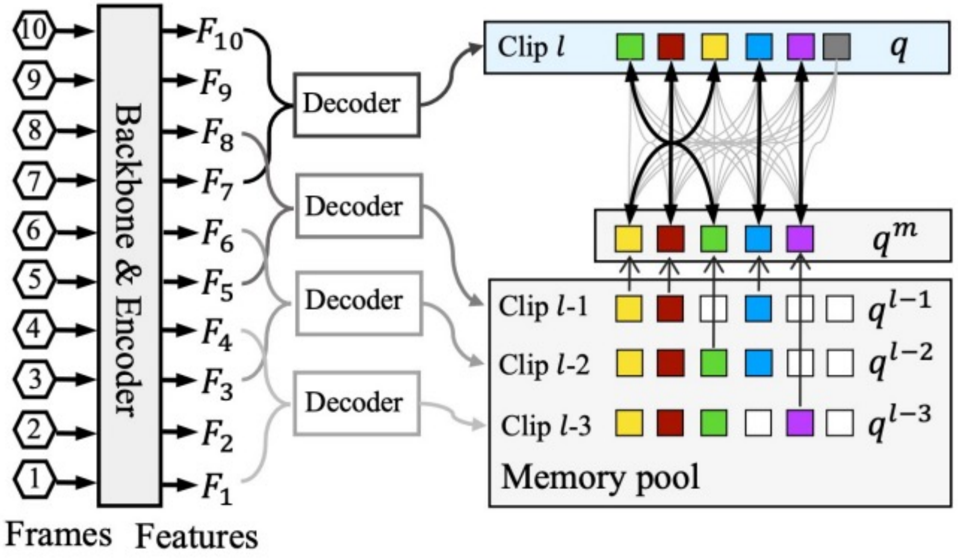



Figure 4. Near-online inference with a clip-by-clip tracker.

| $\beta_1$ | $\beta_2$ | $T_{\text{mem}}$ | mAP | $AP_{50}$ | $AP_{75}$ | $AP_{so}$ | $AP_{mo}$ | $AP_{ho}$ |
|---|---|---|---|---|---|---|---|---|
| 1 | | - | 29.1 | 54.1 | 27.7 | 46.5 | 32.8 | 12.9 |
| | 1 | 10 | 28.3 | 53.4 | 27.1 | 47.1 | 31.3 | 11.6 |
| 1 | 1 | 10 | 30.6 | 57.2 | 28.2 | 49.3 | 35.1 | 13.6 |
| 1 | 1 | 5 | 30.4 | 56.4 | 28.7 | 49.4 | 35.2 | 13.2 |

(d) Tracking. $\beta_1$ and $\beta_2$ control the proportions of mIoU and similarity.

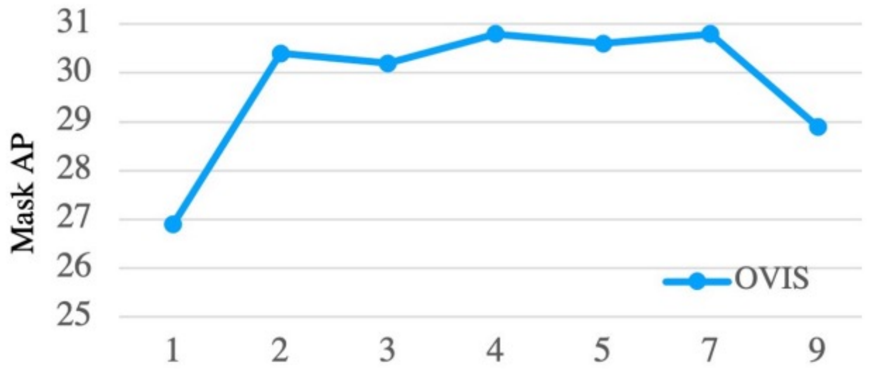



Figure 6. Ablation study on clip length with near-online inference.

# 3.2 Main Results on R50 Backbone

| Type | Methods | YouTube-VIS 2021 AP | YouTube-VIS 2021 $AP_{50}$ | YouTube-VIS 2021 $AP_{75}$ | YouTube-VIS 2021 $AR_1$ | YouTube-VIS 2021 $AR_{10}$ | OVIS AP | OVIS $AP_{50}$ | OVIS $AP_{75}$ | OVIS $AR_1$ | OVIS $AR_{10}$ | FPS | Params |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Per-frame (360p) | MaskTrack [51] | 28.6 | 48.9 | 29.6 | - | - | 10.8 | 25.3 | 8.5 | 7.9 | 14.9 | 20.0 | 58.1M |
| | STMask [26] | 31.1 | 50.4 | 33.5 | 26.9 | 35.6 | 15.4 | 33.9 | 12.5 | 8.9 | 21.4 | 28.0 | - |
| | CrossVIS [52] | 33.3 | 53.8 | 37.0 | 30.1 | 37.6 | 14.9 | 32.7 | 12.1 | 10.3 | 19.8 | 39.8 | 37.5M |
| | InstFormer [24] | 40.8 | 62.4 | 43.7 | 36.1 | 48.1 | 20.0 | 40.7 | 18.1 | 12.0 | 27.1 | - | 44.3M |
| | IDOL [47] | 43.9 | **68.0** | **49.6** | 38.0 | 50.9 | 24.3 | 45.1 | 23.3 | 14.1 | <u>33.2</u> | 30.6 | 43.1M |
| | MinVIS [18] | 44.2 | 66.0 | 48.1 | <u>39.2</u> | <u>51.7</u> | <u>26.3</u> | <u>47.9</u> | <u>25.1</u> | **14.6** | 30.0 | 52.4 | 44.0M |
| Per-clip (360p) | VisTR* [45] | 31.8 | 51.7 | 34.5 | 29.7 | 36.9 | 10.2 | 25.7 | 7.7 | 7.0 | 17.4 | 30.0 | 57.2M |
| | IFC* [19] | 36.6 | 57.9 | 39.3 | - | - | 13.1 | 27.8 | 11.6 | 9.4 | 23.9 | 46.5 | 39.3M |
| | TeViT [53] | 37.9 | 61.2 | 42.1 | 35.1 | 44.6 | 17.4 | 34.9 | 15.0 | 11.2 | 21.8 | 68.9 | 161.8M |
| | SeqFromer* [46] | 40.5 | 62.4 | 43.7 | 36.1 | 48.1 | 15.1 | 31.9 | 13.8 | 10.4 | 27.1 | 72.3 | 49.3M |
| | VITA [17] | **45.7** | <u>67.4</u> | <u>49.5</u> | **40.9** | **53.6** | 19.6 | 41.2 | 17.4 | 11.7 | 26.0 | 33.7 | 57.2M |
| | MDQE (our) | <u>44.5</u> | 67.1 | 48.7 | 37.9 | 49.8 | **29.2** | **55.2** | **27.1** | <u>14.5</u> | **34.2** | 37.8 | 51.4M |
| 720p | IDOL [47] | - | - | - | - | - | 30.2 | 51.3 | 30.0 | 15.0 | 37.5 | - | 43.1M |
| | MDQE (ours) | - | - | - | - | - | **33.0** | **57.4** | **32.2** | **15.4** | **38.4** | 13.5 | 51.4M |
| | MDQE (ours) | - | - | - | - | - | **33.0** | **57.4** | **32.2** | **15.4** | **38.4** | 13.5 | 51.4M |

Table 2. Quantitative performance comparison of VIS methods with ResNet50 backbone on benchmark YouTube-VIS 2021 and OVIS datasets. Note that MinVIS and VITA adopt stronger masked-attention decoder layers proposed in Mask2Former [8]. FPS is computed on YouTube-VIS 2021 valid set, and symbol "-" means the results are not available or applicable. Best in **bold**, second with <u>underline</u>.

# 3.3 Visualization

Frame id: 48 | Frame id: 54 | Frame id: 56 | Frame id: 62

IDOL

MinVIS

MDQE (ours)

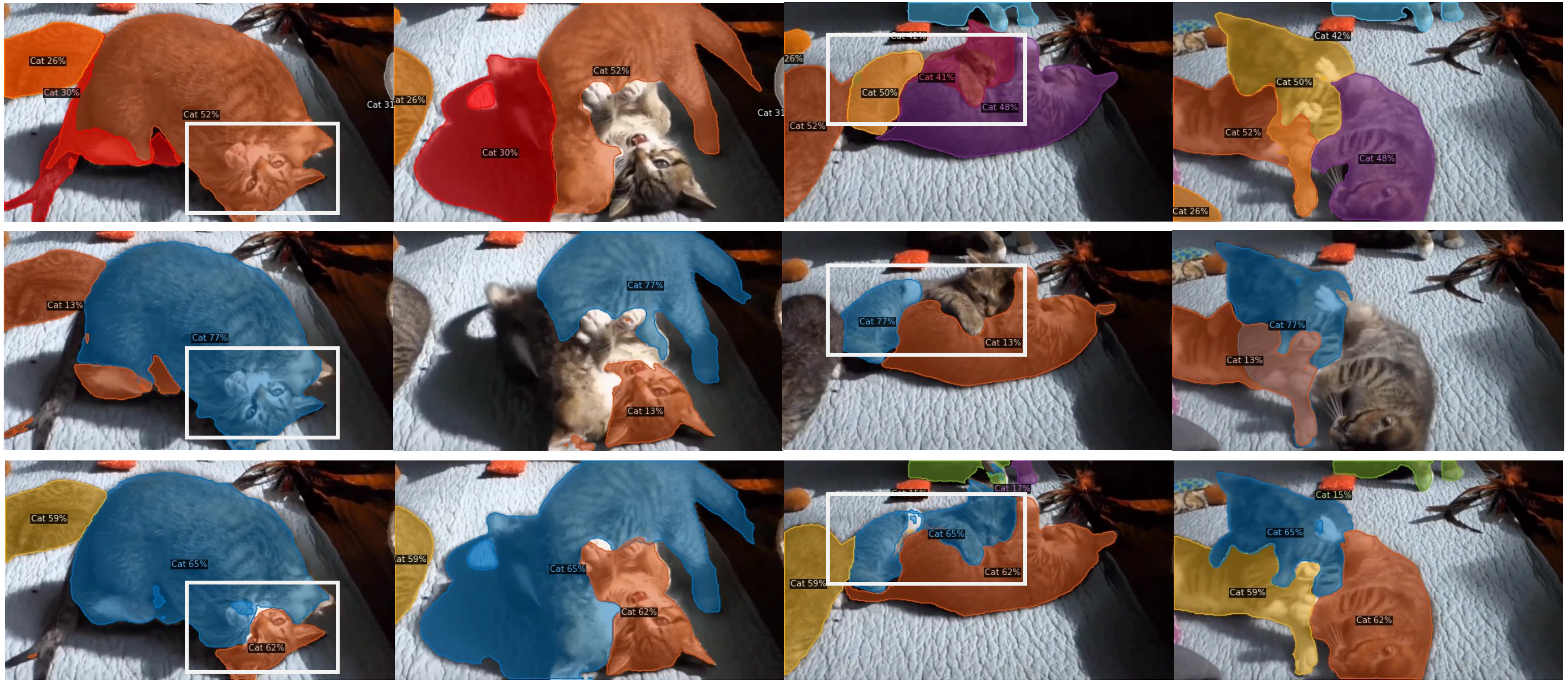
# 3.3 Visualization

Frame id: 36 | Frame id: 56 | Frame id: 76 | Frame id: 94

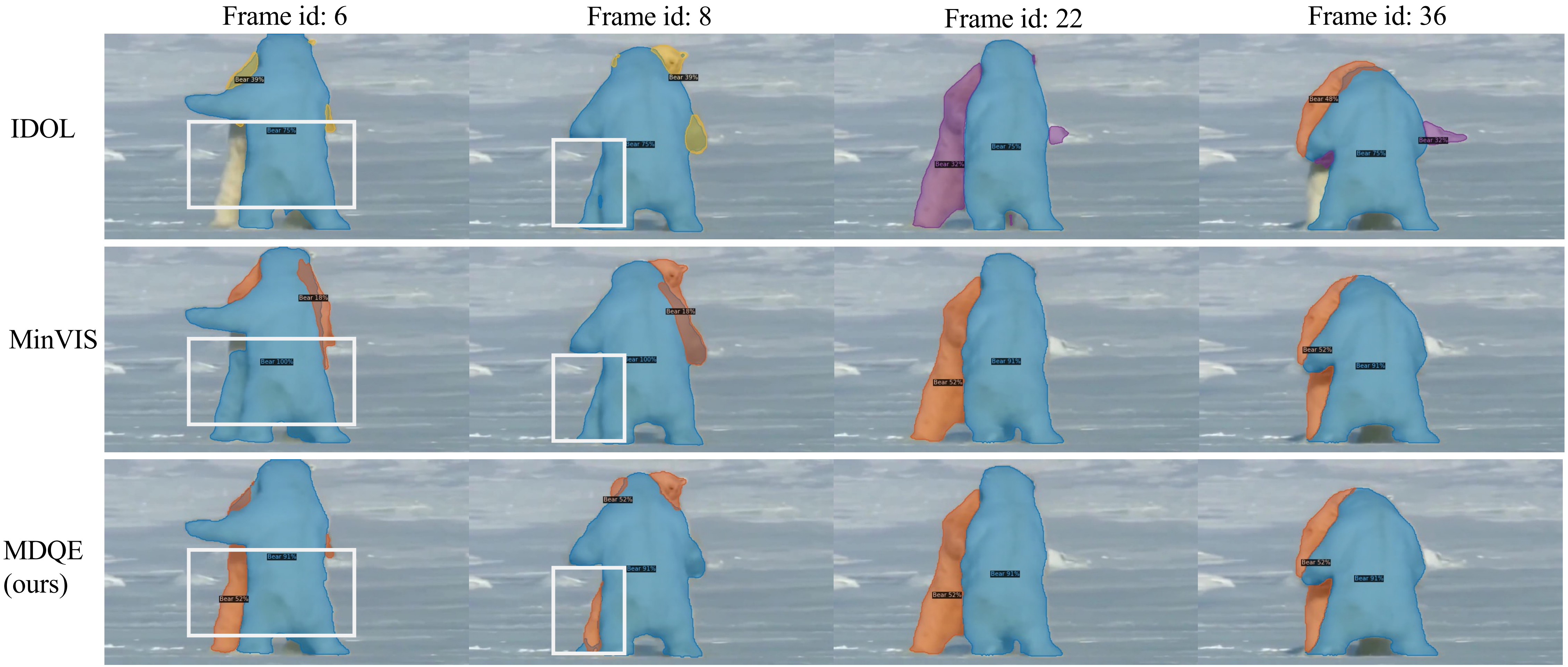
IDOL

MinVIS

MDQE (ours)



# 3.3 Visualization



THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

# Thank you for your attention!

GitHub