MatAnyone:
Stable Video Matting with Consistent Memory Propagation

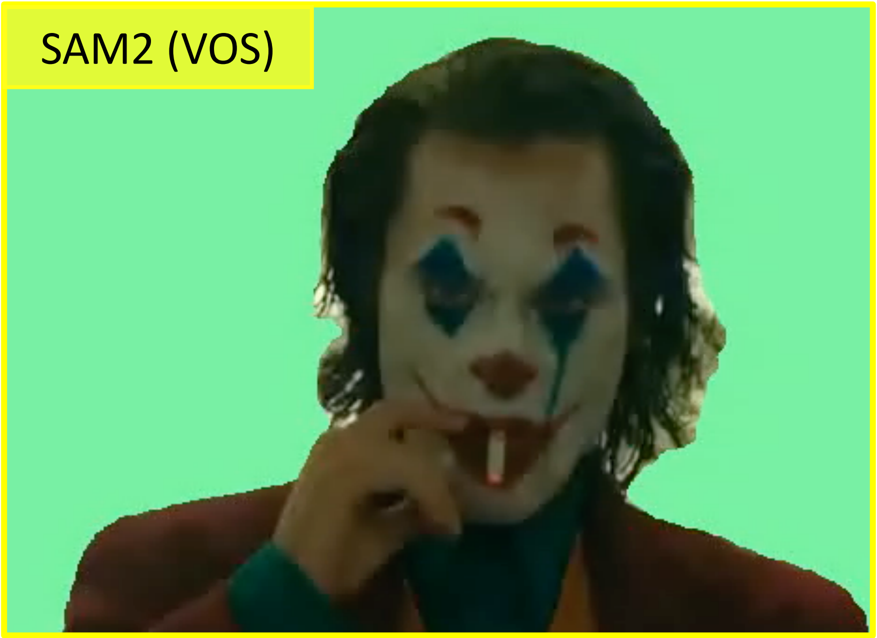
Peiqing Yang[1], Shangchen Zhou[1], Jixin Zhao[1], Qingyi Tao[2], Chen Change Loy[1]
[1]S-Lab, Nanyang Technological University, [2]SenseTime Research, Singapore

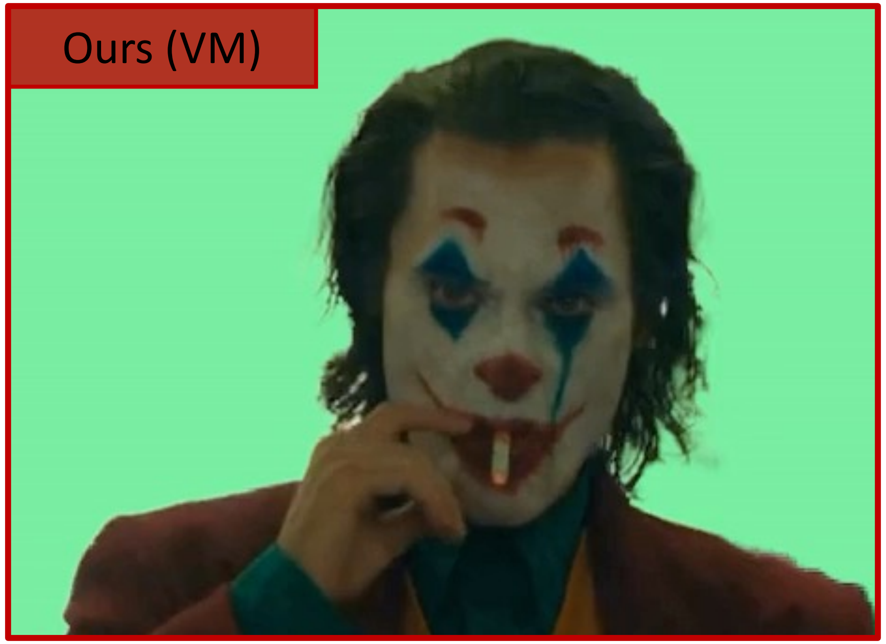
🎸 **CVPR 2025** 🔥 **1K GitHub Stars**

MMLab@NTU | S-Lab, Nanyang Technological University

What is Video Matting and What are the Applications?

Input Video

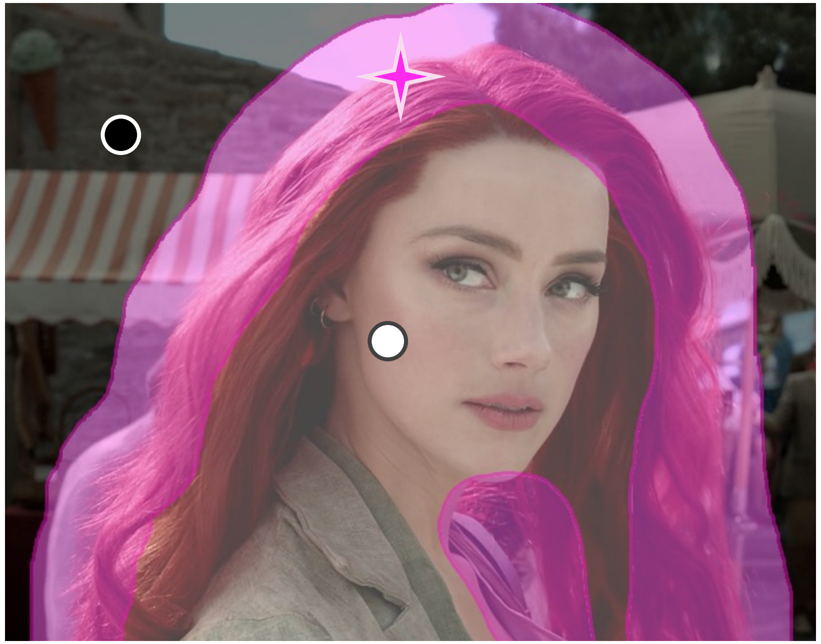




Video Segmentation *vs.* Video Matting

- Video Matting (VM) poses additional challenges compared to Video Segmentation (VOS)

- VM requires:
  - [Core Areas] Accurate semantic detection
  - [Boundary Area] High-quality detail extraction



Core Areas Boundary Area

Applications: Real-world Use Cases

Virtual Background

Background Replacement

Visual Effects (VFX) Editing

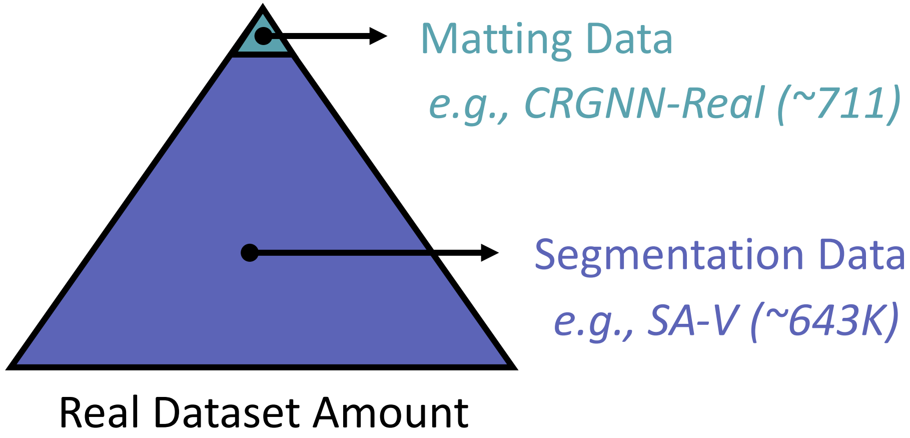
https://support.zoom.com/hc/zh/article?id=zm_kb&sysparm_article=KB0060398
https://www.youtube.com/watch?v=-tQCqvBhM6o
https://www.youtube.com/watch?v=gyeif8yaHhM

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

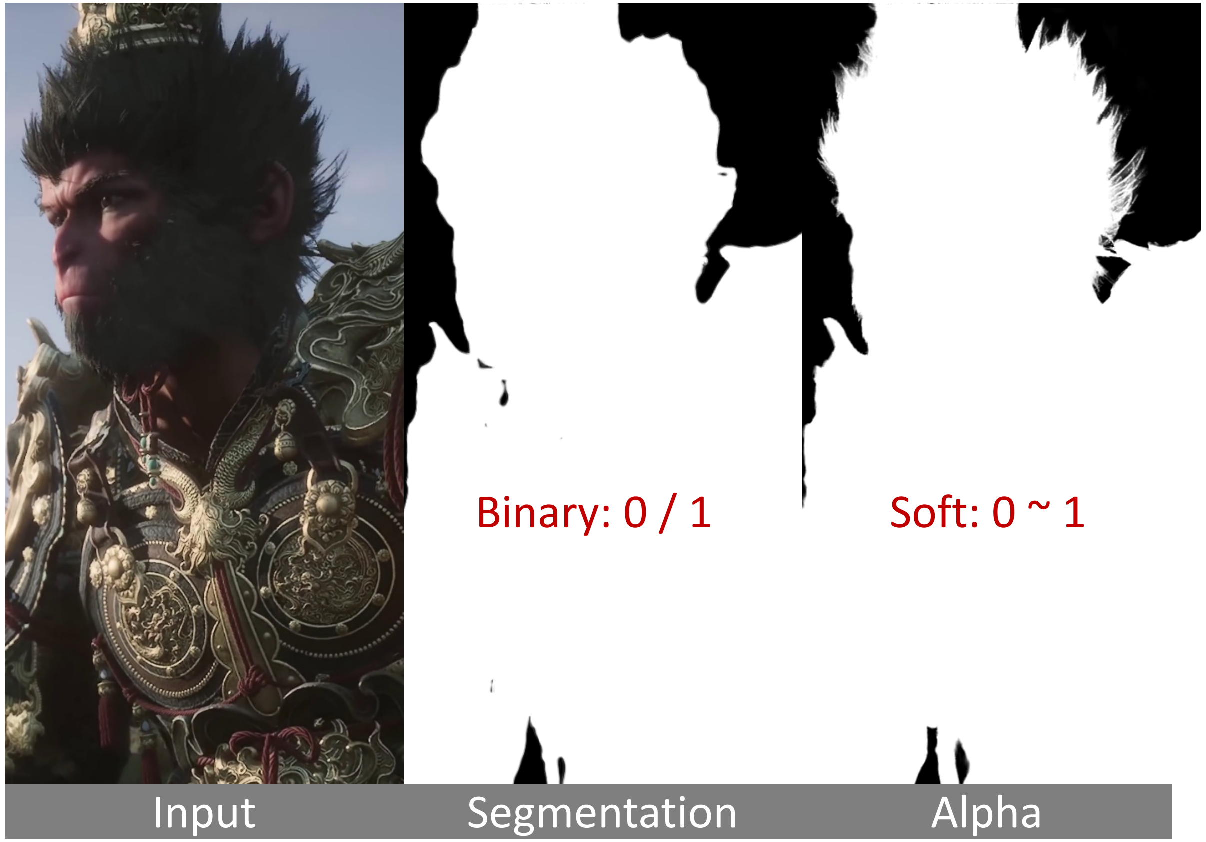
S-LAB FOR ADVANCED INTELLIGENCE

What makes Video Matting even more Challenging?

Challenge: Data



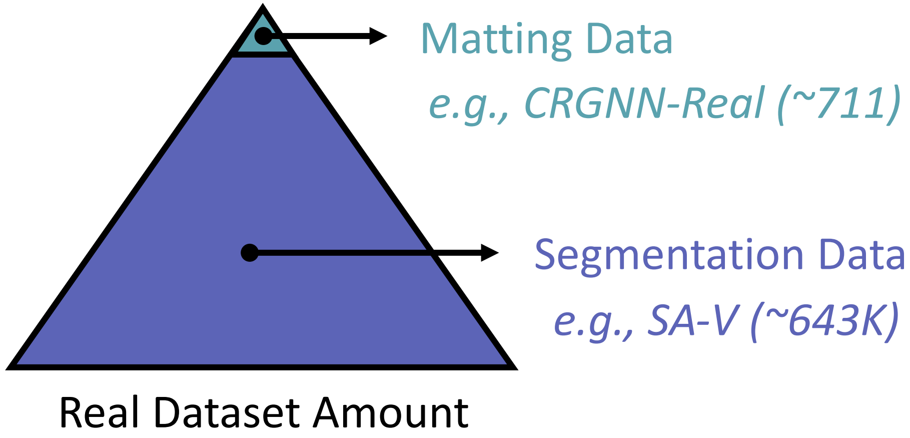
- Lack of *large-scale* real data with *alpha* masks



- Extremely high annotation costs
- If image is still possible, video is nearly impossible

Challenge: Data

- Lack of *large-scale* real data with *alpha* masks



Input w/ Given Seg Mask | Matting Output (MaGGIe)

☞ MaGGIe Segmentation prior broken

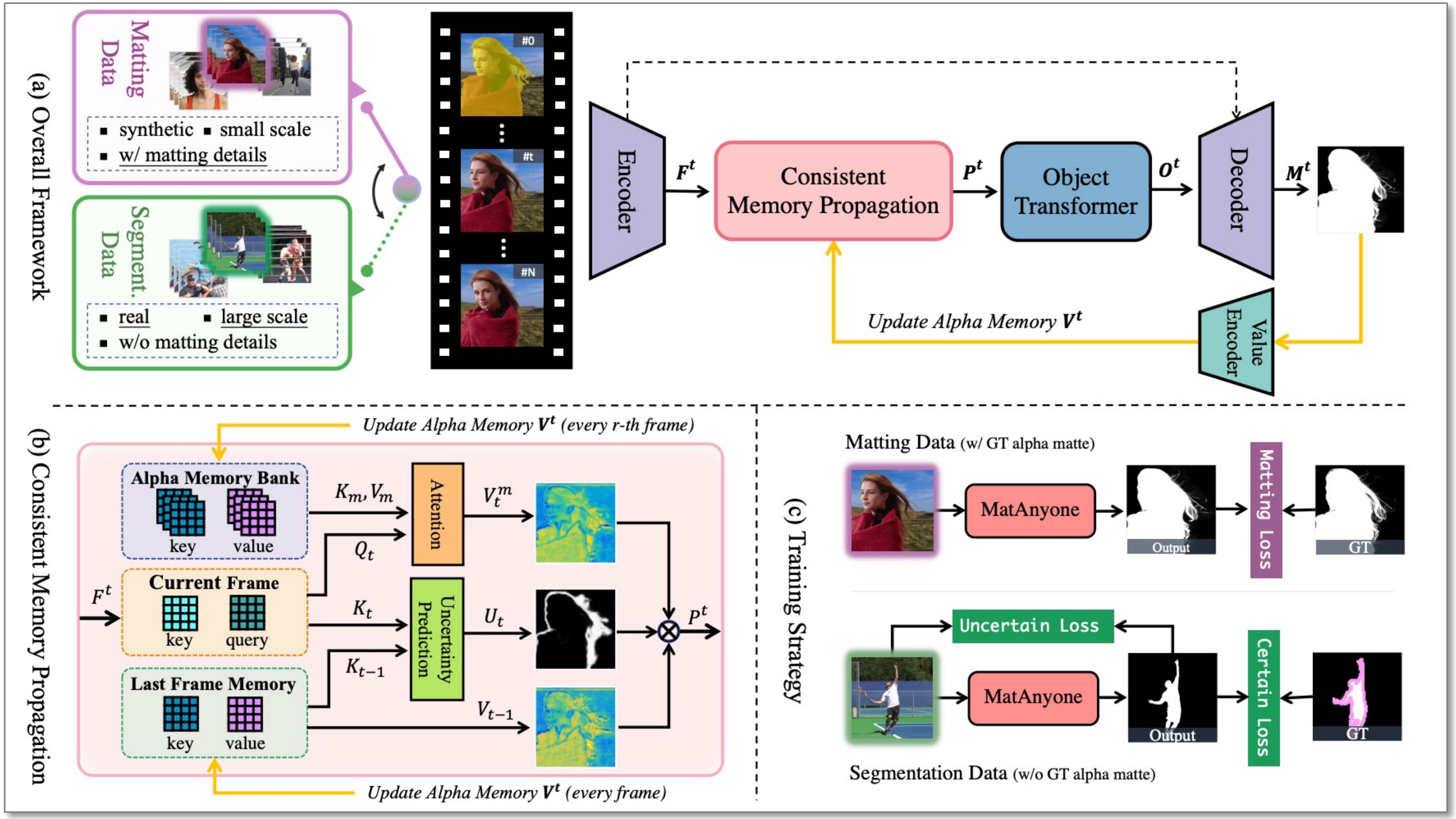
- ❖ Currently, only synthetic data available
- ❖ Distribution Gap: Harms real-world performance

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

S-LAB FOR ADVANCED INTELLIGENCE

What are our <u>Key Designs</u> to tackle the challenges?

Our Framework

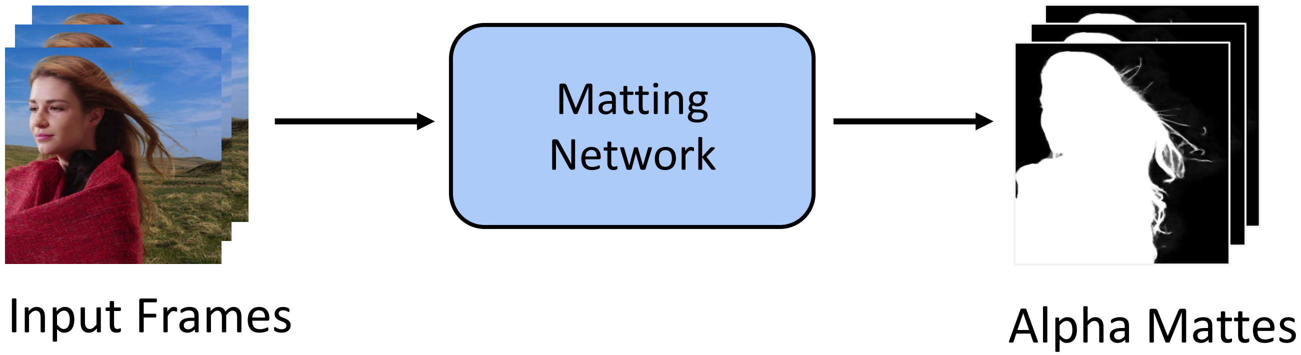


Key designs in:

- Network
- Training Strategy
- Data

Current Methods

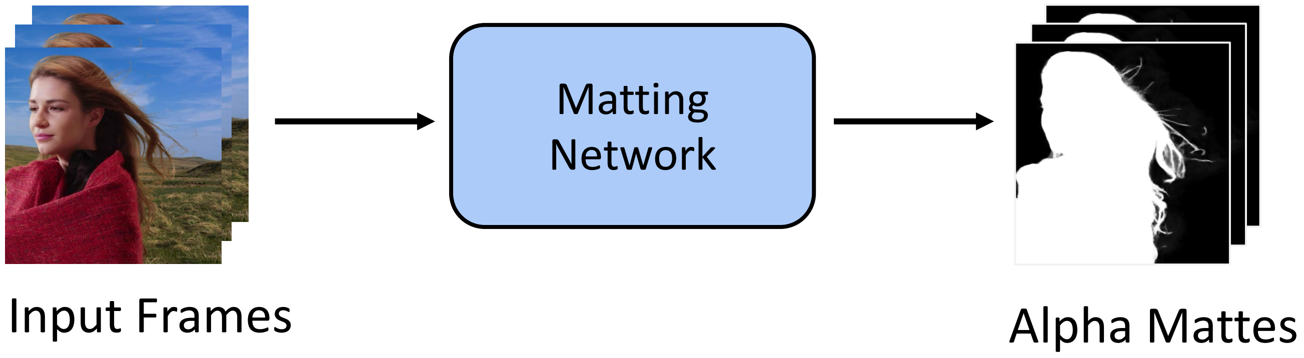
Auxiliary-free Methods (MODNet, RVM)



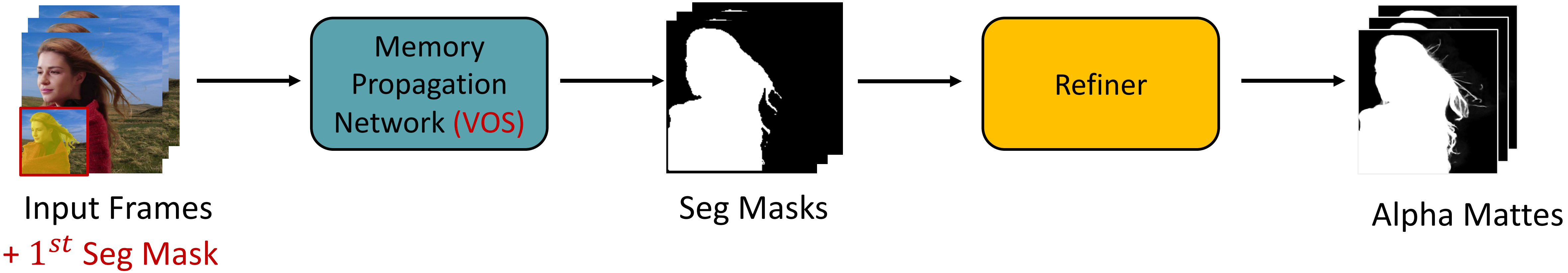
A Trimap-Free Portrait Matting Solution in Real Time (AAAI 2022)
Robust High-Resolution Video Matting with Temporal Guidance (WACV 2022)

Current Methods

Auxiliary-free Methods (MODNet, RVM)



Mask-guided Methods (AdaM, MaGGIe)

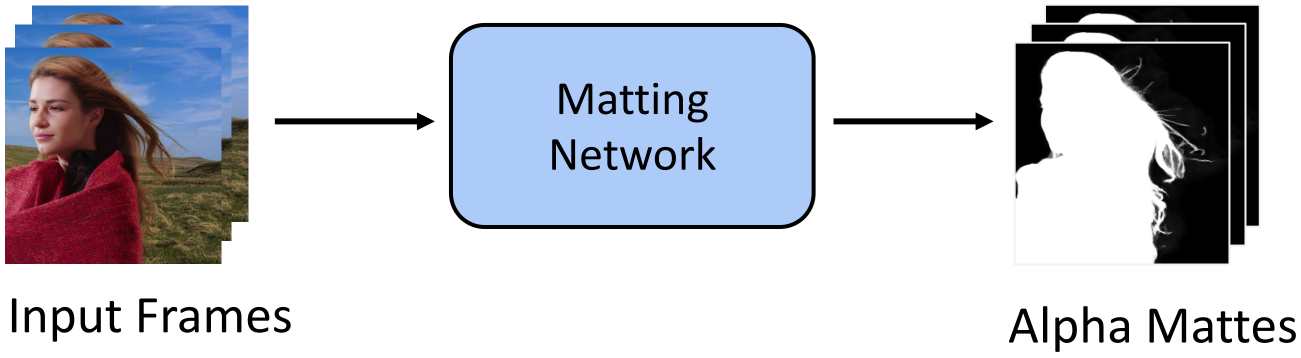


Adaptive Human Matting for Dynamic Videos (CVPR 2023)

MaGGIe: Mask Guided Gradual Human Instance Matting (CVPR 2024)

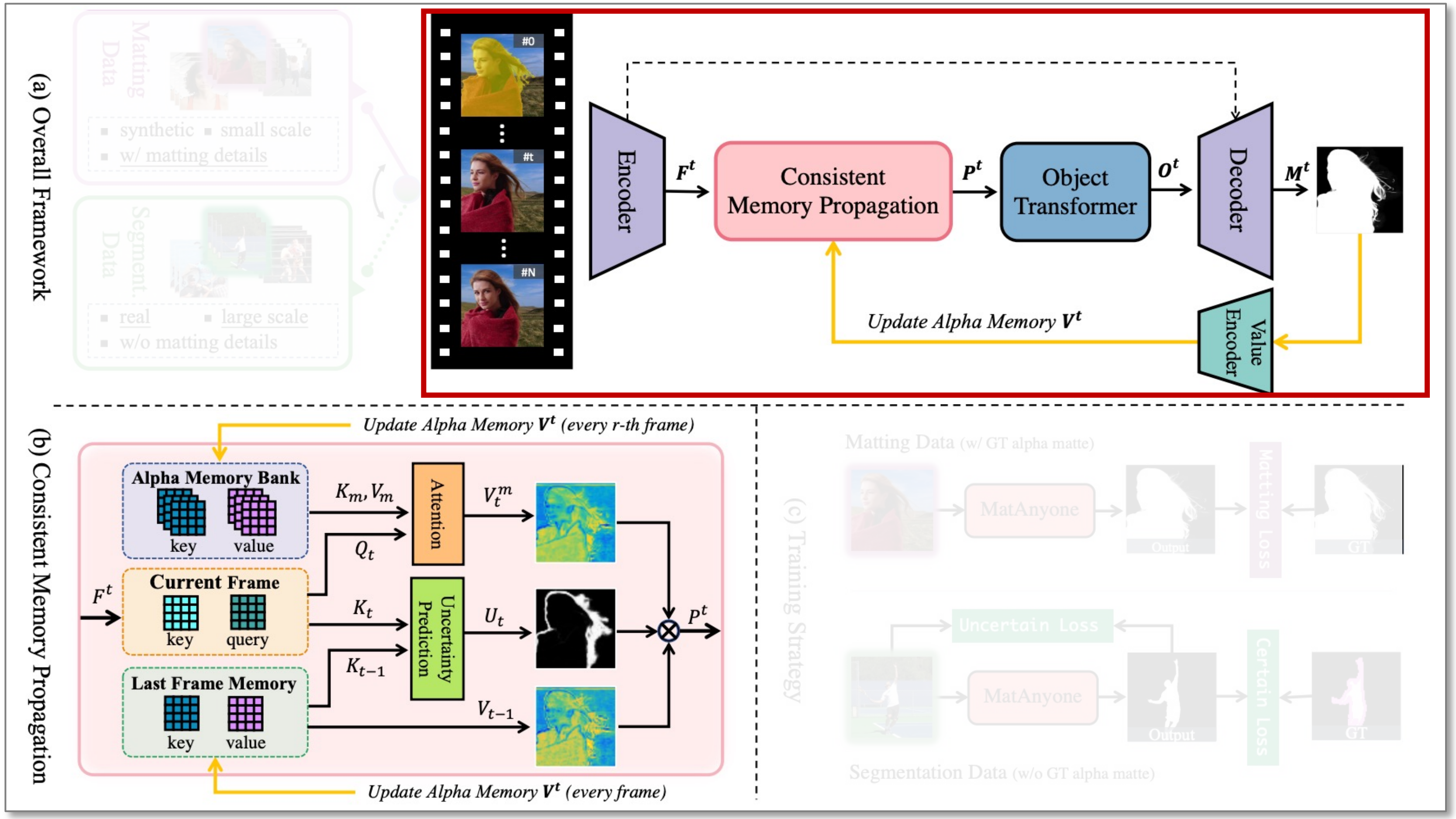
Current Methods

Auxiliary-free Methods (MODNet, RVM)



Mask-guided Methods (Ours)

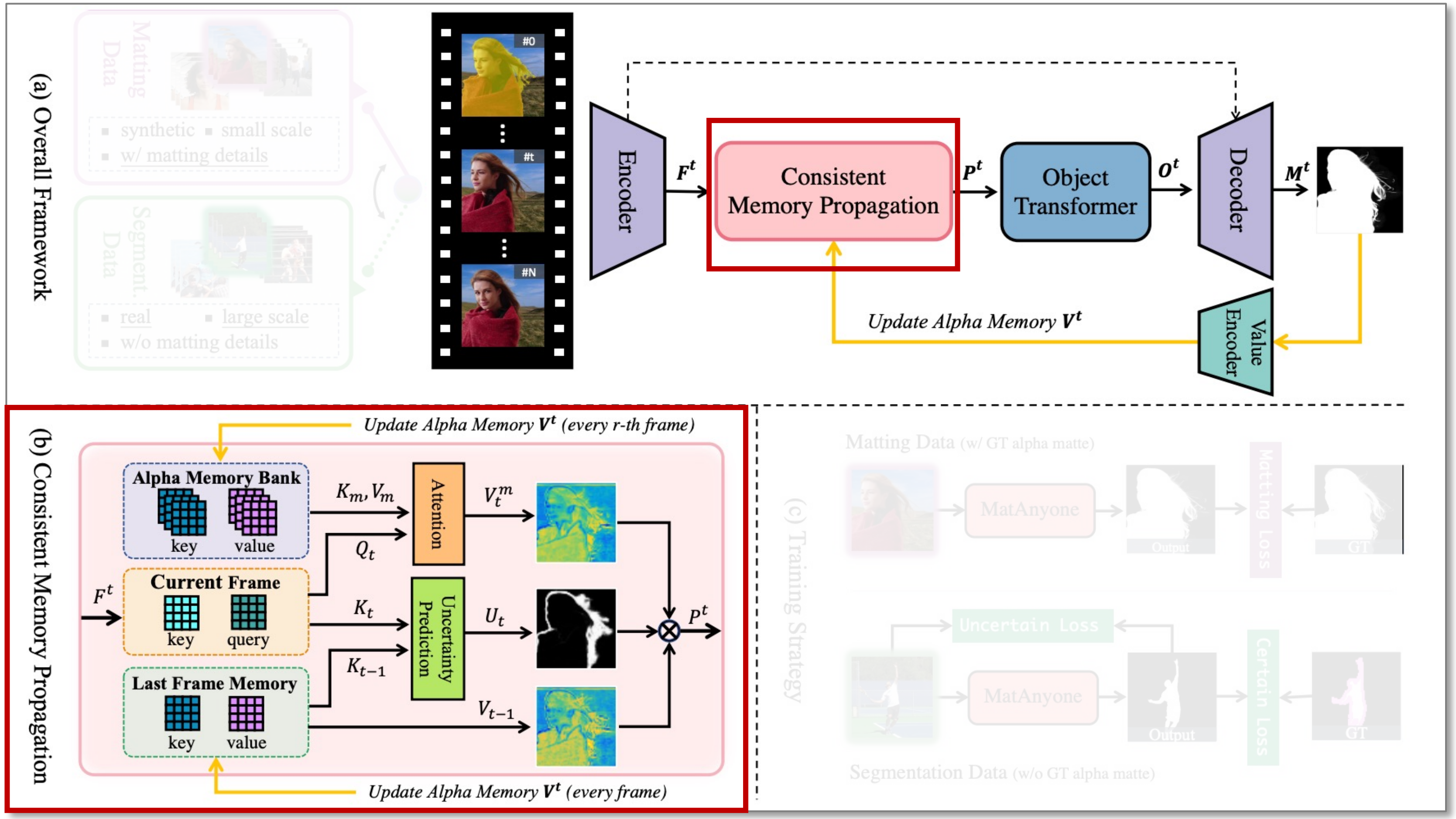
Network Design



(1) Mask-guided VM:

Given first-frame *segmentation mask*

Network Design



(1) Mask-guided VM:

Given first-frame *segmentation mask*

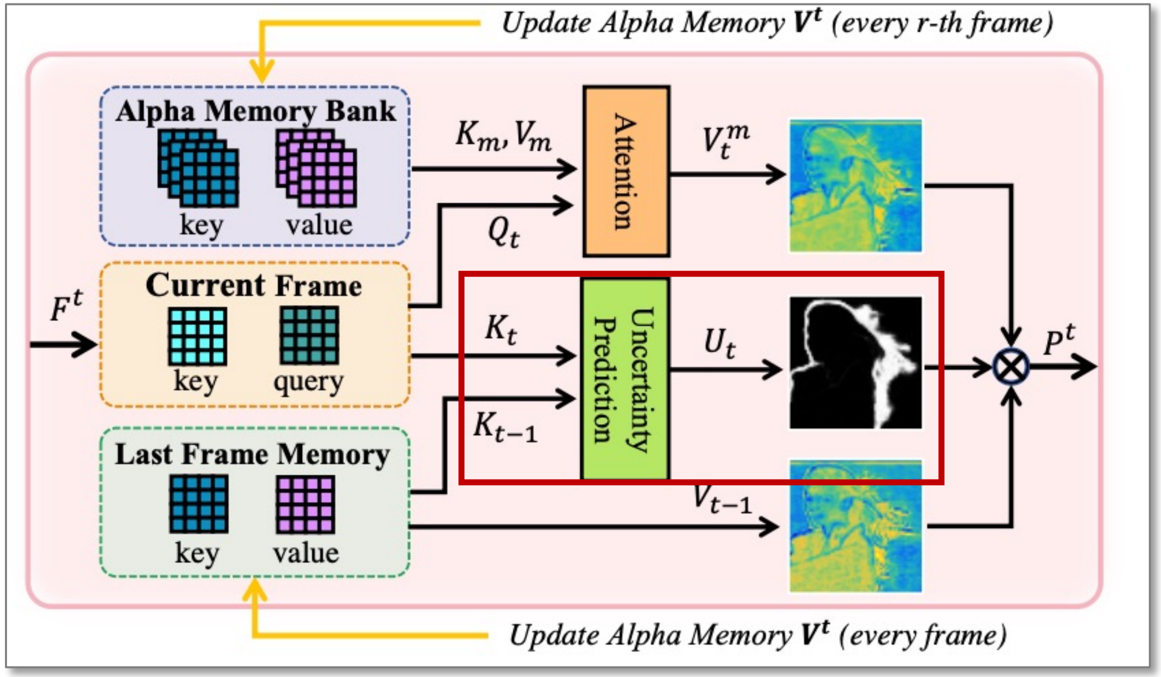
(2) Consistent Memory Propagation:

Region-adaptive memory fusion

Consistent Memory Propagation (CMP)

Region-adaptive memory fusion:

- ”Change” probability: $U_t \in [0,1]$



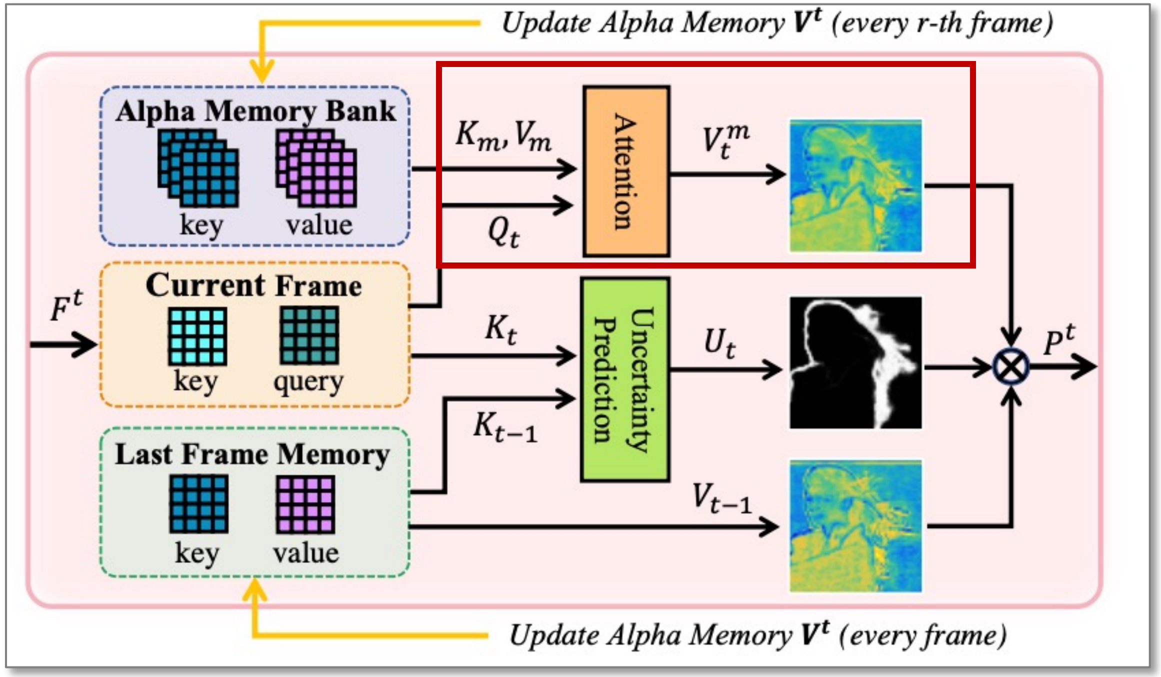
$$P_t = V_t^m * U_t + V_{t-1} * (1 - U_t)$$

Consistent Memory Propagation (CMP)

Region-adaptive memory fusion:

- ”Change” probability: $U_t \in [0, 1]$
- “Large-change” region:

  Mainly from <u>memory bank</u> (V_t^m)



$$P_t = V_t^m * U_t + V_{t-1} * (1 - U_t)$$

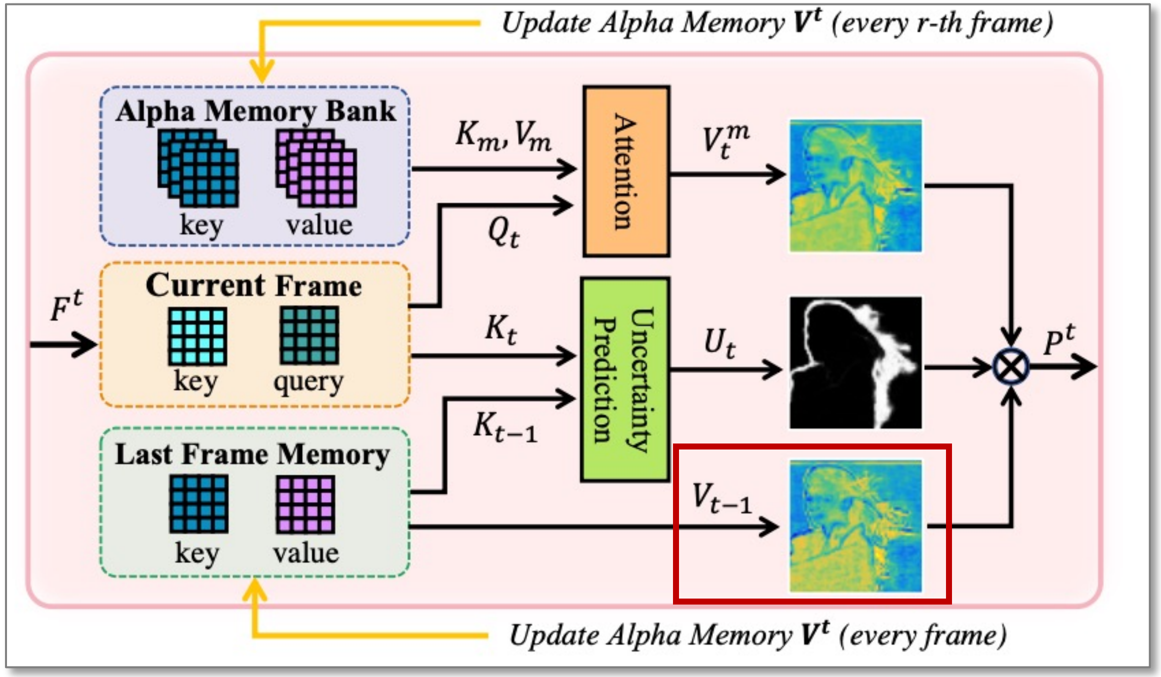
Consistent Memory Propagation (CMP)

Region-adaptive memory fusion:

- ”Change” probability: $U_t \in [0, 1]$
- “Large-change” region:

  Mainly from memory bank (V_t^m)
- “small-change” region:

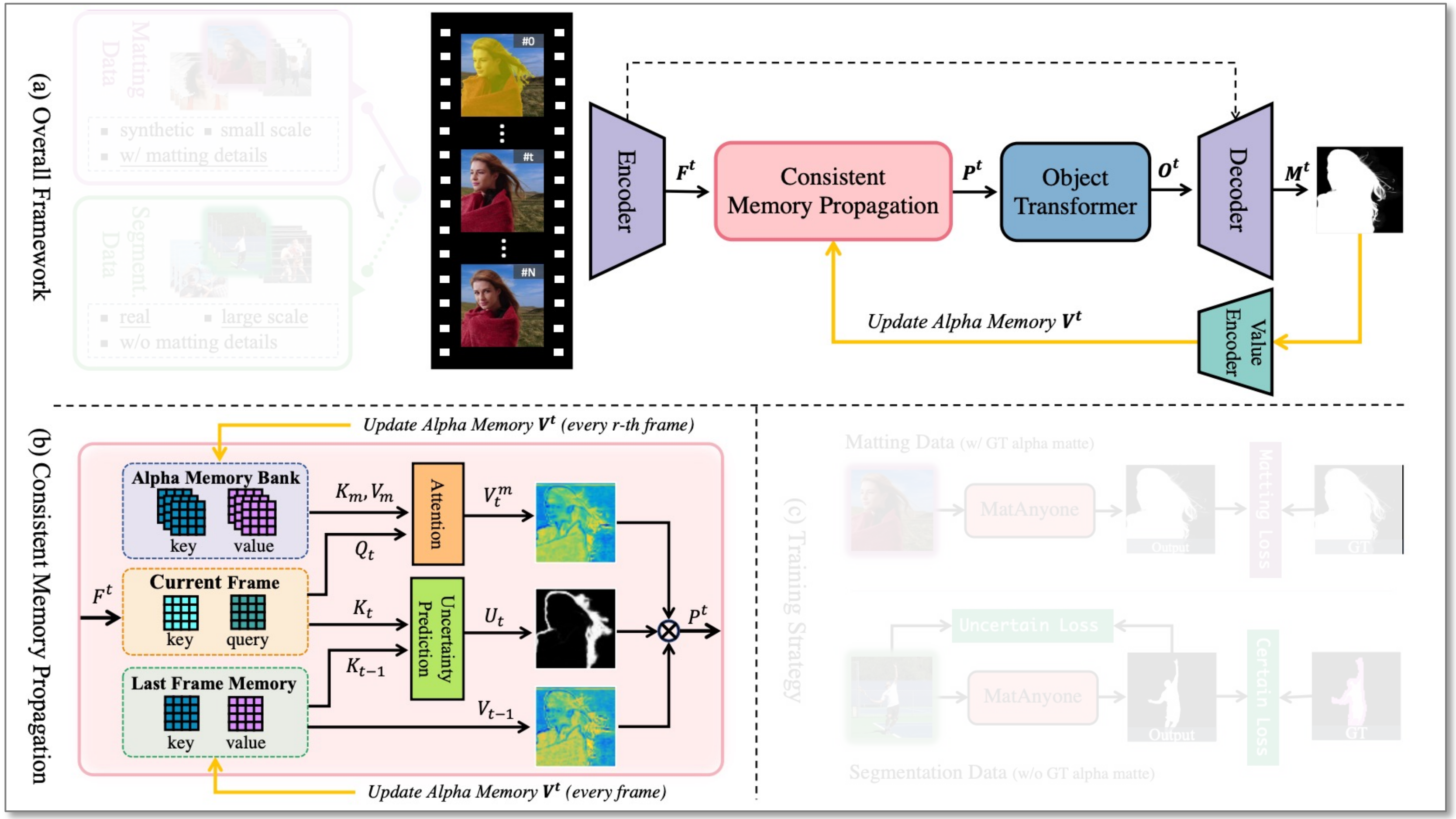
  Mainly from <u>last frame</u> (V_{t-1})



$$P_t = V_t^m * U_t + V_{t-1} * (1 - U_t)$$

Ablation: Effectiveness of CMP

Network Design



(1) Mask-guided VM:

Given first-frame *segmentation mask*

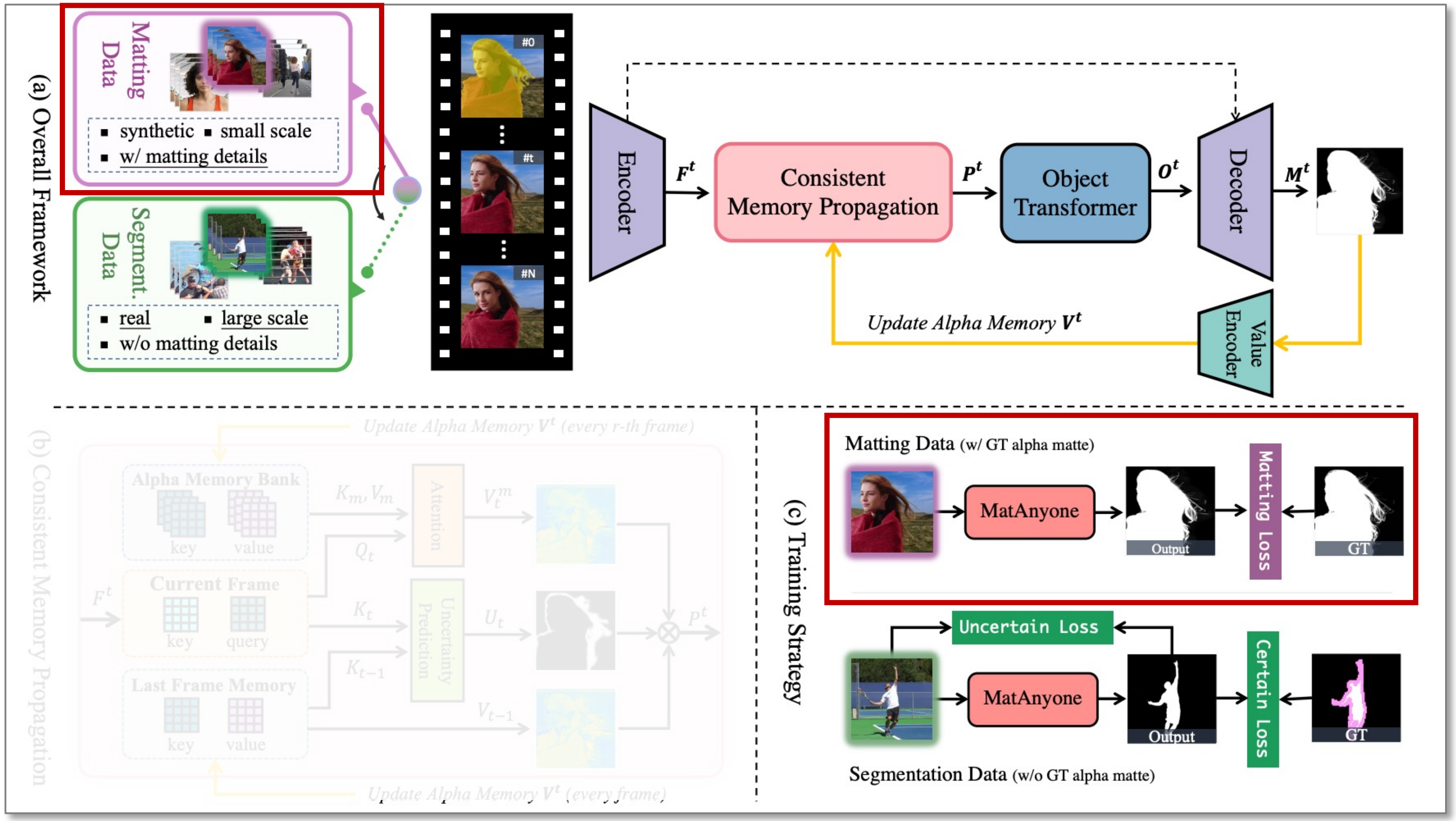
(2) Consistent Memory Propagation:

Region-adaptive memory fusion

(3) Recurrent Refinement:

To reach the image-matting level

Training Strategy Design



(1) Matting Data:

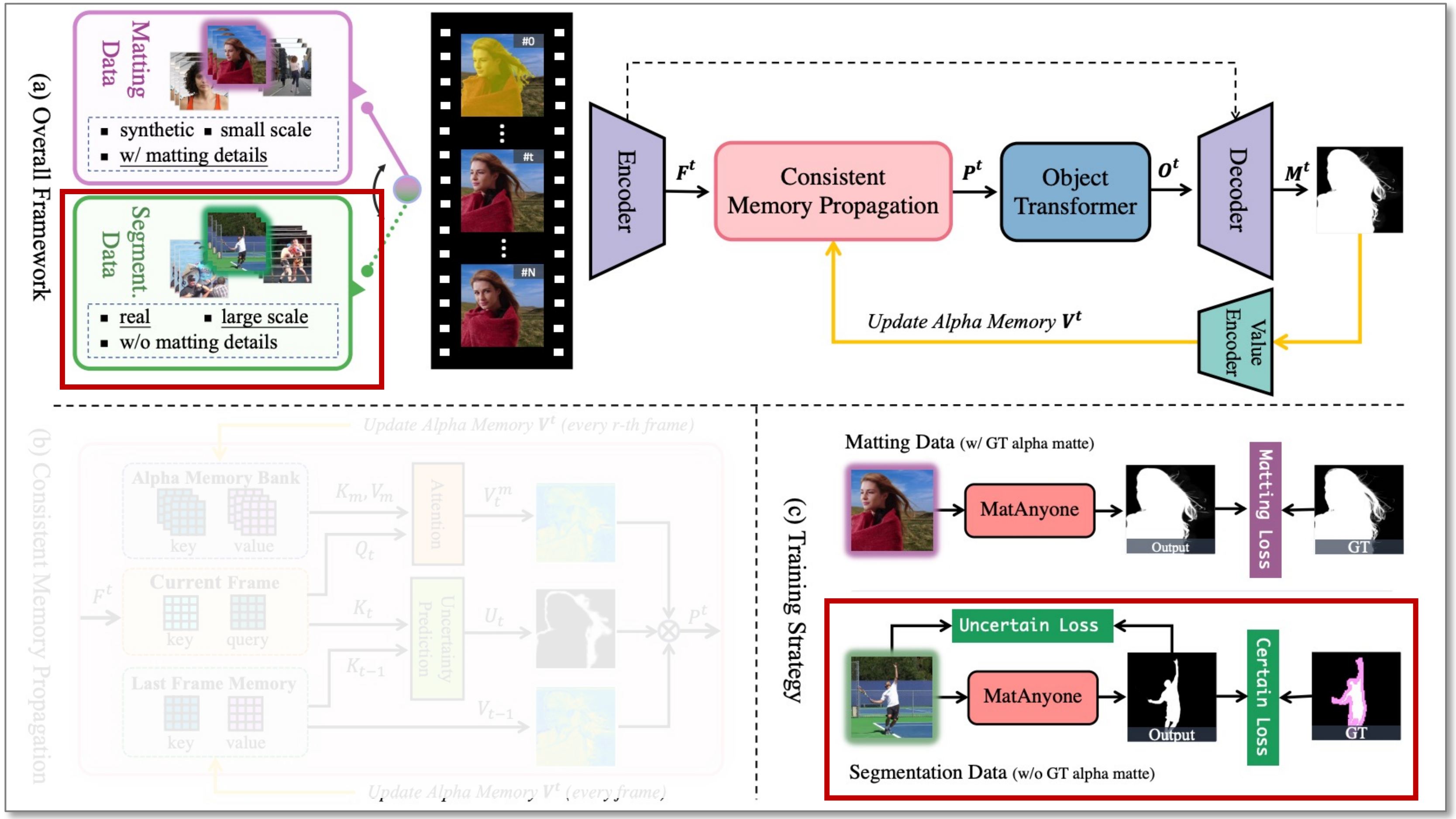
$$\mathcal{L}_{l1} = \|M_t - M_t^{GT}\|_1,$$

$$\mathcal{L}_{lap} = \sum_{s=1}^{5} \frac{2^{s-1}}{5} \|L_{pyr}^s(M_t) - L_{pyr}^s(M_t^{GT})\|_1$$

$$\mathcal{L}_{tc} = \|\frac{\mathrm{d}M_t}{\mathrm{d}t} - \frac{\mathrm{d}M_t^{GT}}{\mathrm{d}t}\|_2$$

$$\boxed{\mathcal{L}^{mat} = \mathcal{L}_{l1} + 5\mathcal{L}_{lap} + \mathcal{L}_{tc}}$$

Training Strategy Design



(2) Segmentation Data:

(Region-specific loss)

$$\mathcal{L} = \mathcal{L}_{certain} + \mathcal{L}_{uncertain}$$

Core region (w/ label) → $L1\ loss$

Boundary region (w/o label) → ?

How to supervise without GT alpha labels?

- DDC loss: supervise with input image ONLY

$$\mathcal{L}_{DDC} = \frac{1}{N}\sum_{i}^{N}\sum_{j}|\alpha_i - \alpha_j - \|\boldsymbol{I}_i - \boldsymbol{I}_j\|_2|$$
$$j \in \text{argtopk}\{-\|\boldsymbol{I}_i - \boldsymbol{I}_j\|_2\}$$

- We propose scaled DDC loss to *relax* originally strict assumptions:

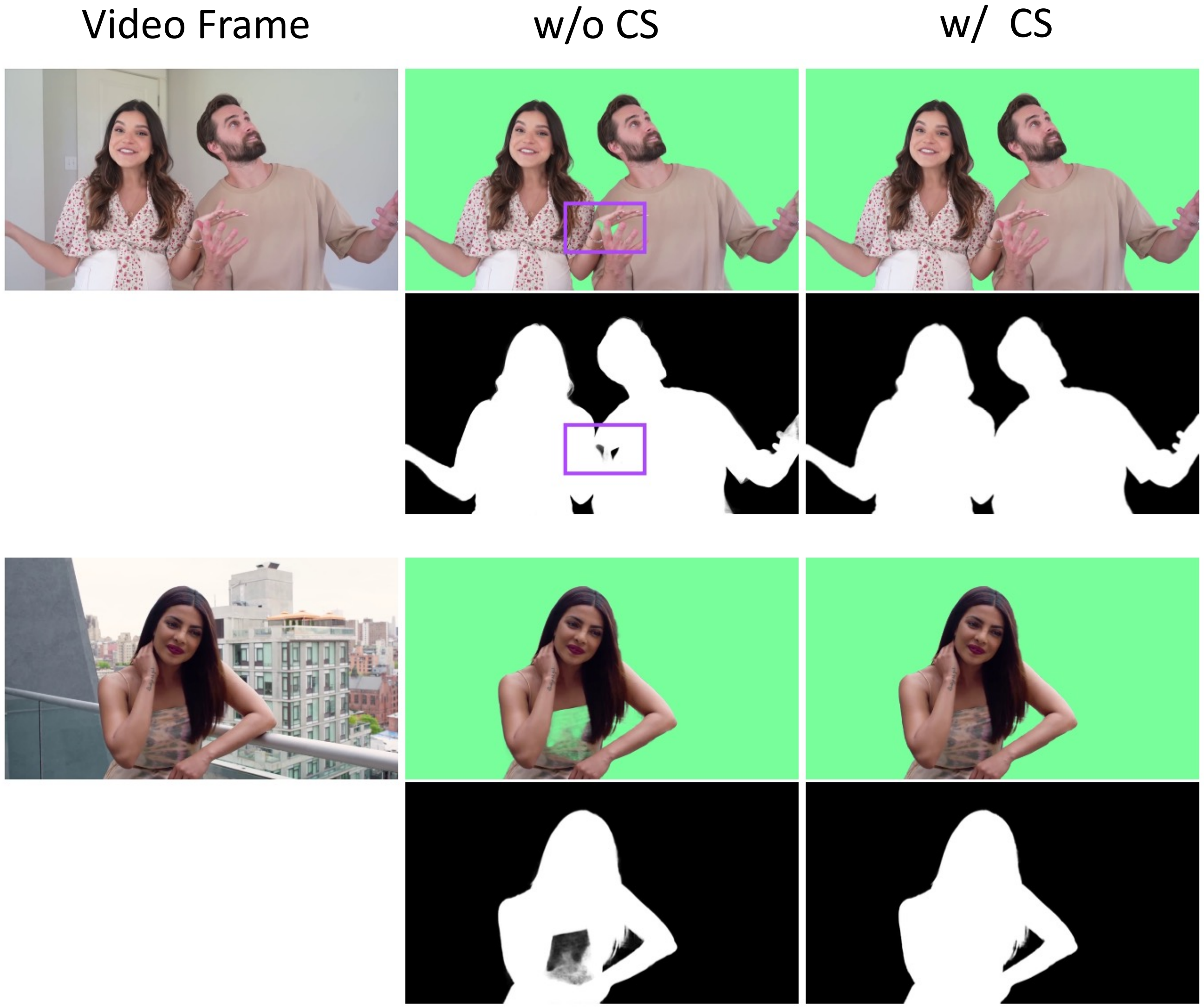
$$\mathcal{L}_{boundary} = \frac{1}{N}\sum_{i}^{N}\sum_{j}|(\alpha_i - \alpha_j)(\mathbf{F} - \mathbf{B}) - \|\boldsymbol{I}_i - \boldsymbol{I}_j\|_2|$$
$$j \in \text{argtopk}\{-\|\boldsymbol{I}_i - \boldsymbol{I}_j\|_2\}$$

- We call such strategy of using segmentation data as core supervision (CS):

$$\mathcal{L}^{cs} = \mathcal{L}_{core} + 1.5\mathcal{L}_{boundary}$$

Video Frames | DDC Loss | Scaled DDC loss

Ablation: Effectiveness of Core Supervision (CS)



- Previous strategy: obvious semantics error due to the *weak* supervision from real segmentation data
- Our strategy: largely improves semantics accuracy thanks to the *stronger* supervision enabled with core supervision loss

Data Design

Training Data

Datesets	VideoMatte240K (old train) [32]	**VM800** (new train)	VideoMatte (old test) [32]	**YouTubeMatte** (new test)
#Foregrounds	475	826	5	32
Sources	-	Storyblocks, Envato Elements, Motion Array	-	YouTube
Harmonized	-	-	x	✓

Processing Pipeline

Keylight
- Screen Color: pixel value of upper left corner
- Screen Matte:
  - Clip Black: 20
  - Clip White: 80

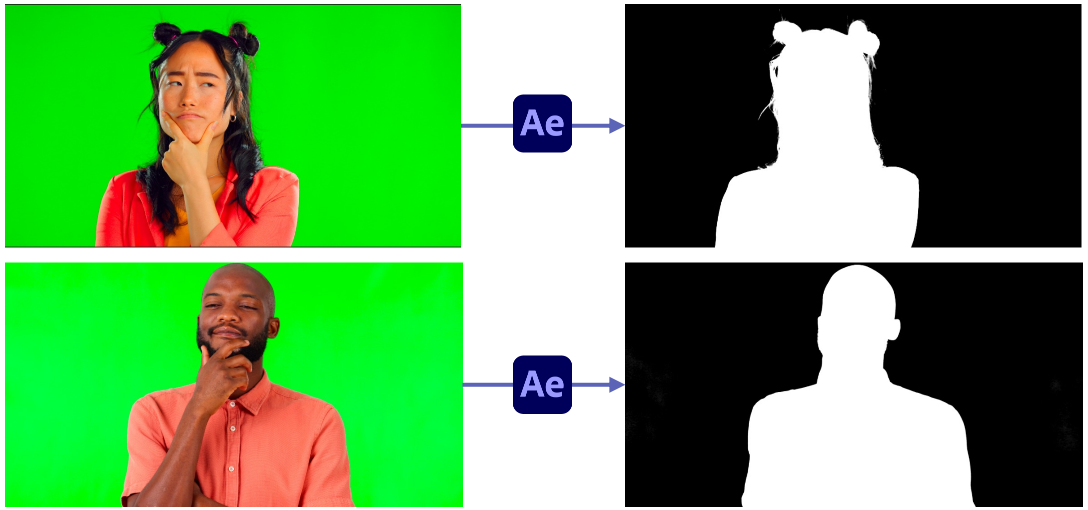
Key Cleaner
- radius: 1
- reduce chatter: check

Advanced Spill Supressor

→ **Save as QuickTime (.mov) : RGB + Alpha**

→ JSX

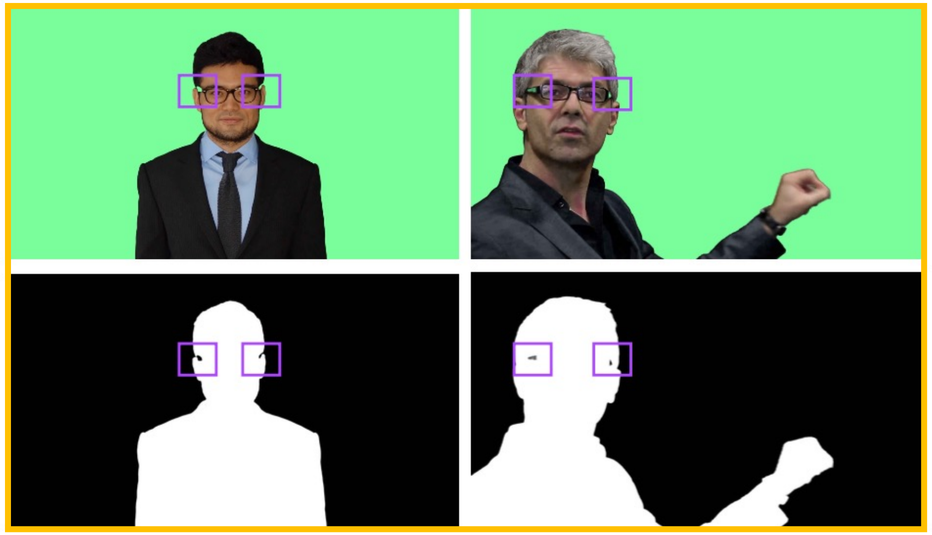
pipeline.jsx

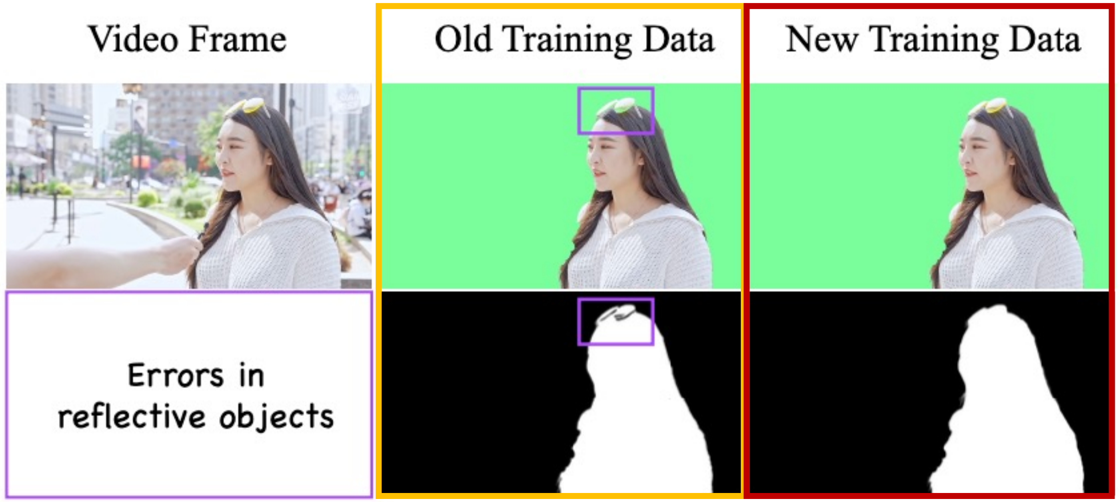


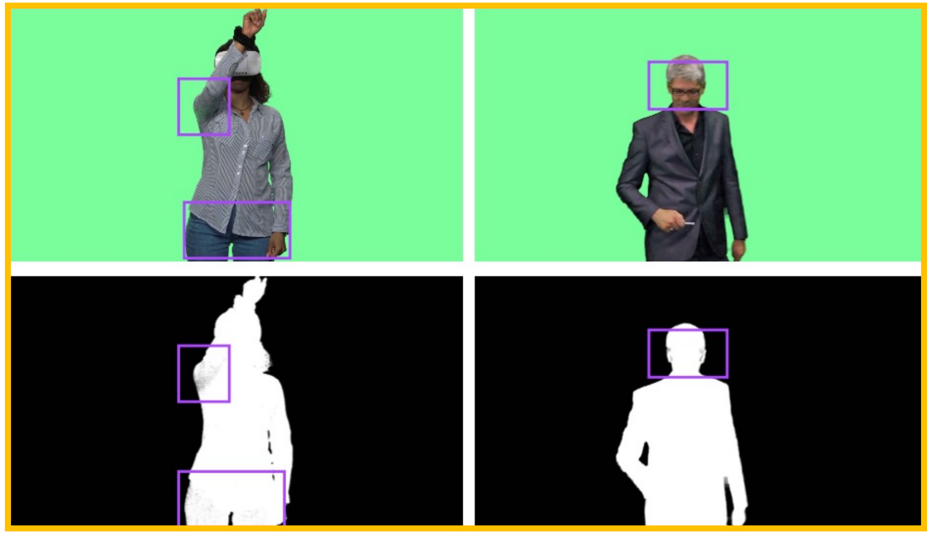
Green-screen Footage

Alpha mattes

Ablation: Enhancement from New Training Data







Data Design

Testing Benchmark

Datesets	VideoMatte240K (old train) [32]	**VM800** (new train)	VideoMatte (old test) [32]	**YouTubeMatte** (new test)
#Foregrounds	475	826	5	32
Sources	-	Storyblocks, Envato Elements, Motion Array	-	YouTube
Harmonized	-	-	x	✓

Harmonization when compositing

Before

After

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

S-LAB FOR ADVANCED INTELLIGENCE

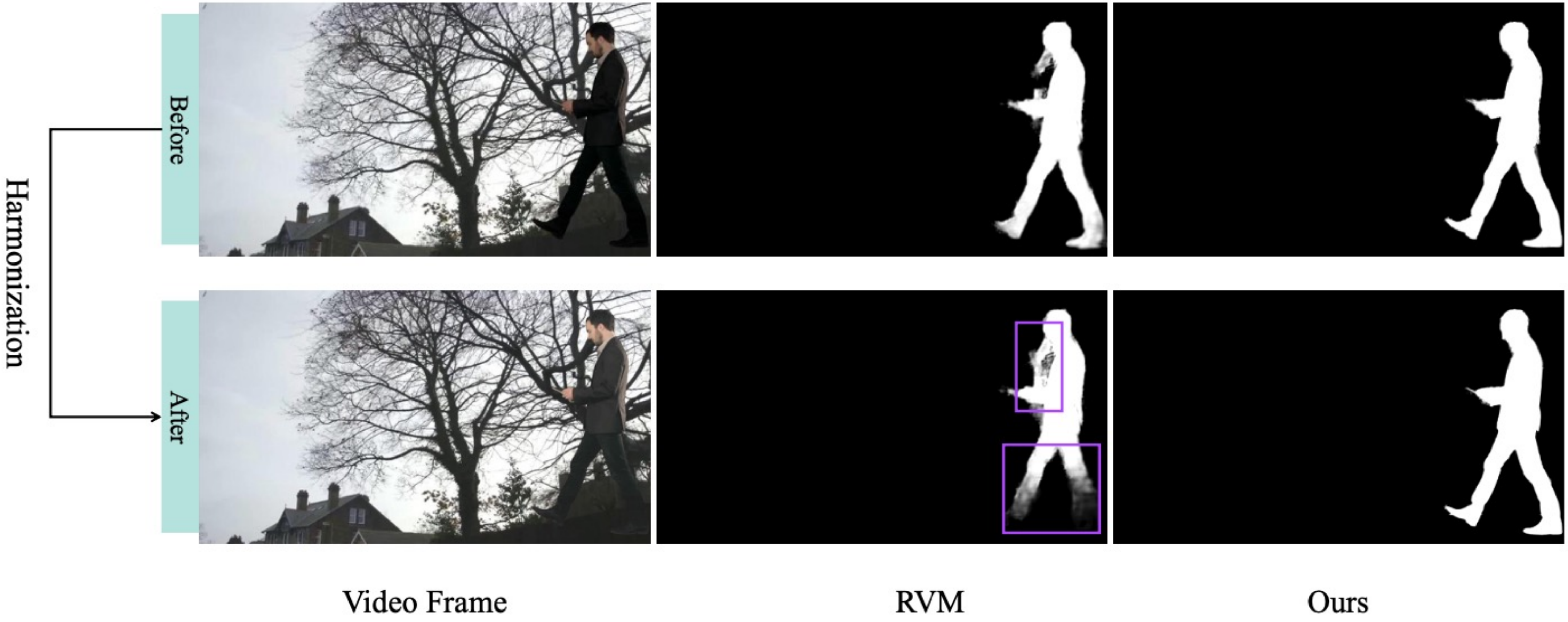
How does our model Perform?

Experiment Results – Synthetic Dataset

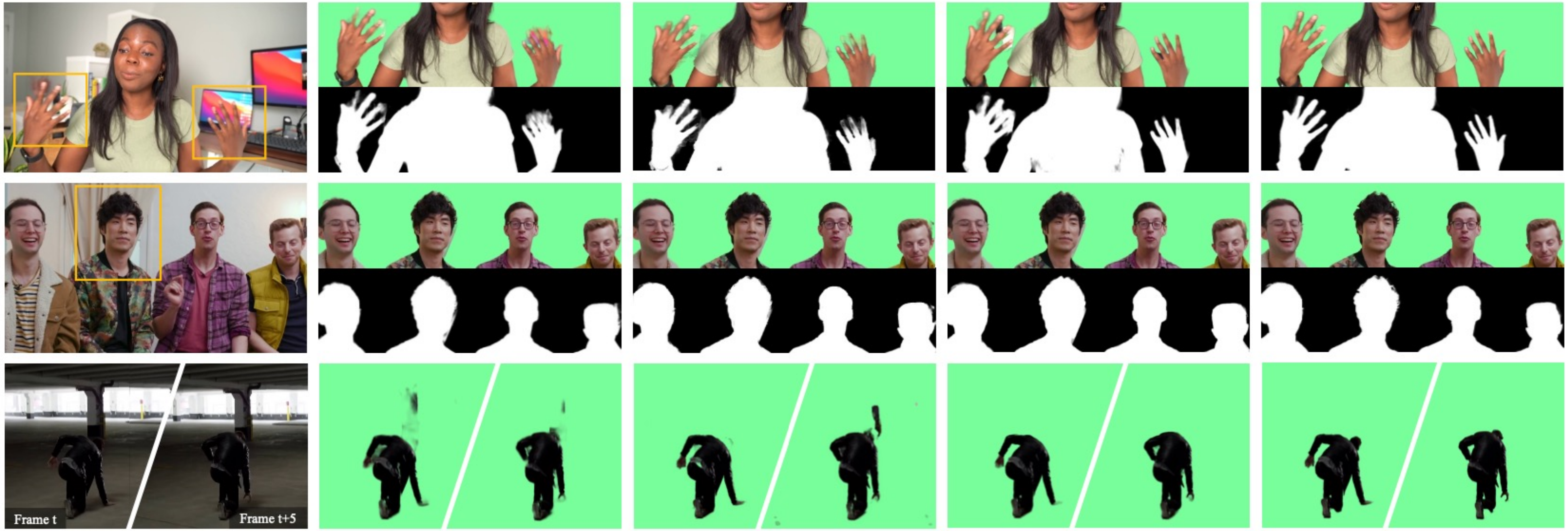
Metrics	Auxiliary-free (AF) Methods			Mask-guided Methods			
	MODNet [24]	RVM [33]	RVM-Large [33]	AdaM [31]	FTP-VM [20]	MaGGIe† [22]	**Ours**
VideoMatte (512 × 288)							
MAD↓	9.41	6.08	5.32	5.30	6.13	5.49	5.15
MSE↓	4.30	1.47	0.62	0.78	1.31	0.60	0.93
Grad↓	1.89	0.88	0.59	0.72	1.14	0.57	0.67
dtSSD↓	2.23	1.36	1.24	1.33	1.60	1.39	1.18
Conn↓	0.81	0.41	0.30	0.30	0.41	0.31	0.26
VideoMatte (1920 × 1080)							
MAD↓	11.13	6.57	5.81	4.42	8.00	4.42	4.24
MSE↓	5.54	1.93	0.97	0.39	3.24	0.40	0.33
Grad↓	15.30	10.55	9.65	5.12	23.75	4.03	4.00
dtSSD↓	3.08	1.90	1.78	1.39	2.37	1.31	1.19
YoutubeMatte (512 × 288)							
MAD↓	19.37	4.08	3.36	-	3.08	3.54	2.72
MSE↓	16.21	1.97	1.04	-	1.29	1.23	1.01
Grad↓	2.05	1.34	1.03	-	1.16	1.10	0.97
dtSSD↓	2.79	1.81	1.62	-	1.83	1.88	1.60
Conn↓	2.68	0.60	0.50	-	0.41	0.49	0.39
YoutubeMatte (1920 × 1080)							
MAD↓	15.29	4.37	3.58	-	6.49	2.37	1.99
MSE↓	12.68	2.25	1.23	-	4.58	0.98	0.71
Grad↓	8.42	15.1	12.97	-	29.78	7.69	8.91
dtSSD↓	2.74	2.28	2.04	-	2.41	1.77	1.65

- Best MAD:
  - Spatial Accuracy
- Best dtSSD:
  - Temporal Stability
- Best Conn:
  - Visual Quality

Experiment Results – Synthetic Dataset

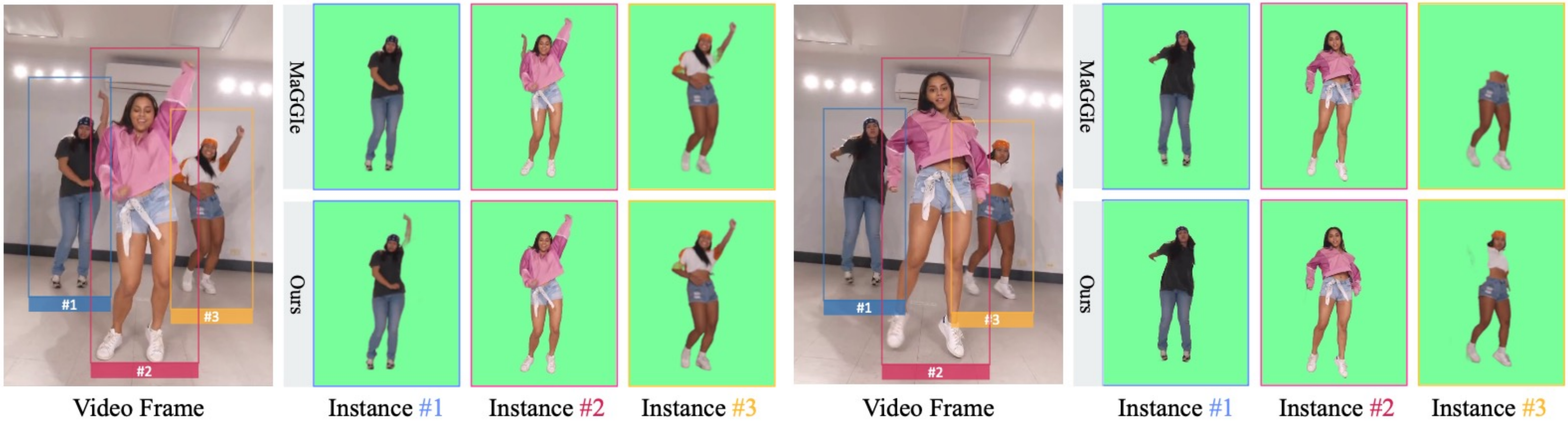


Real Results - General Video Matting



Video Frame RVM FTP-VM MaGGIe Ours

Real Results - Instance Video Matting

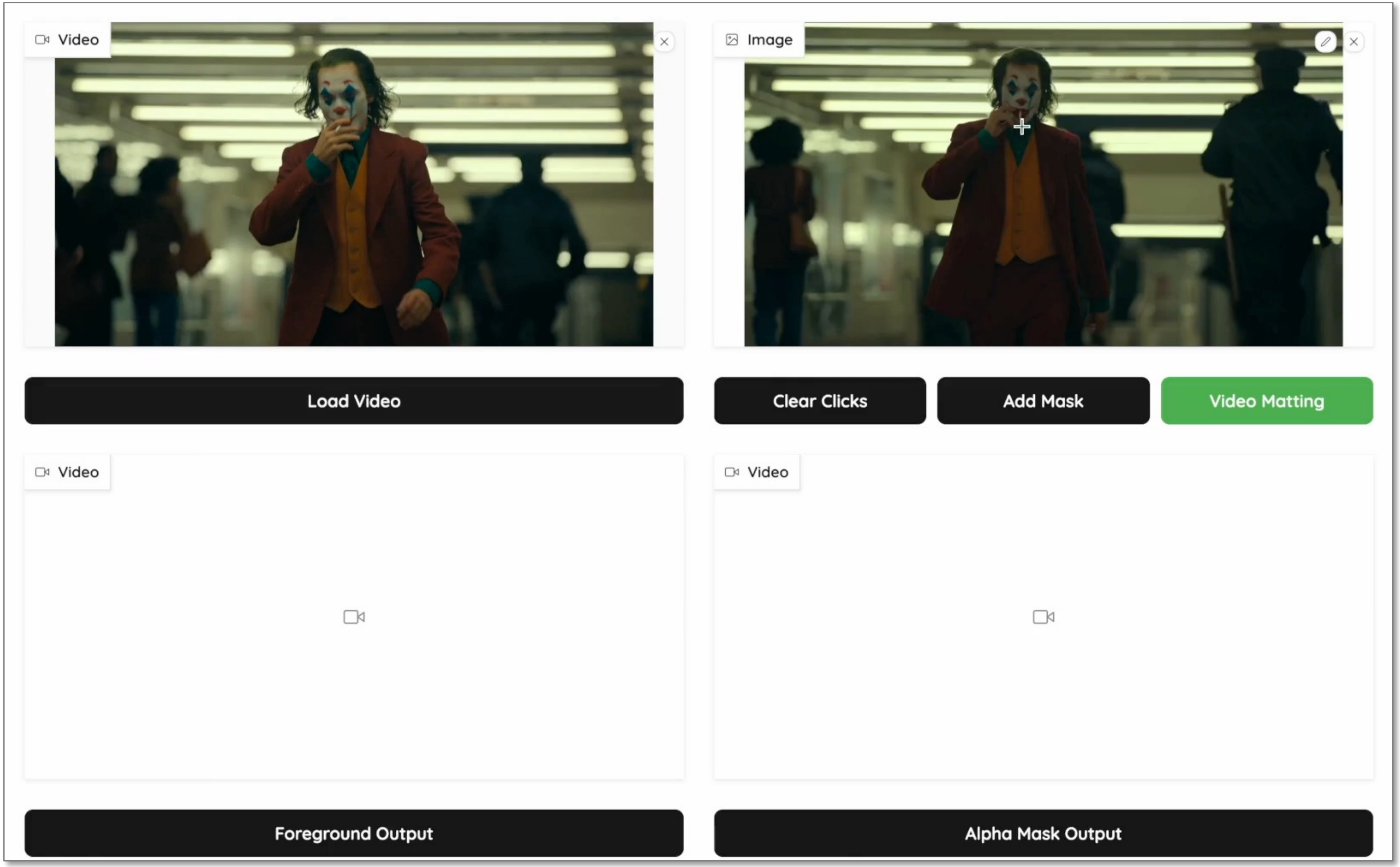


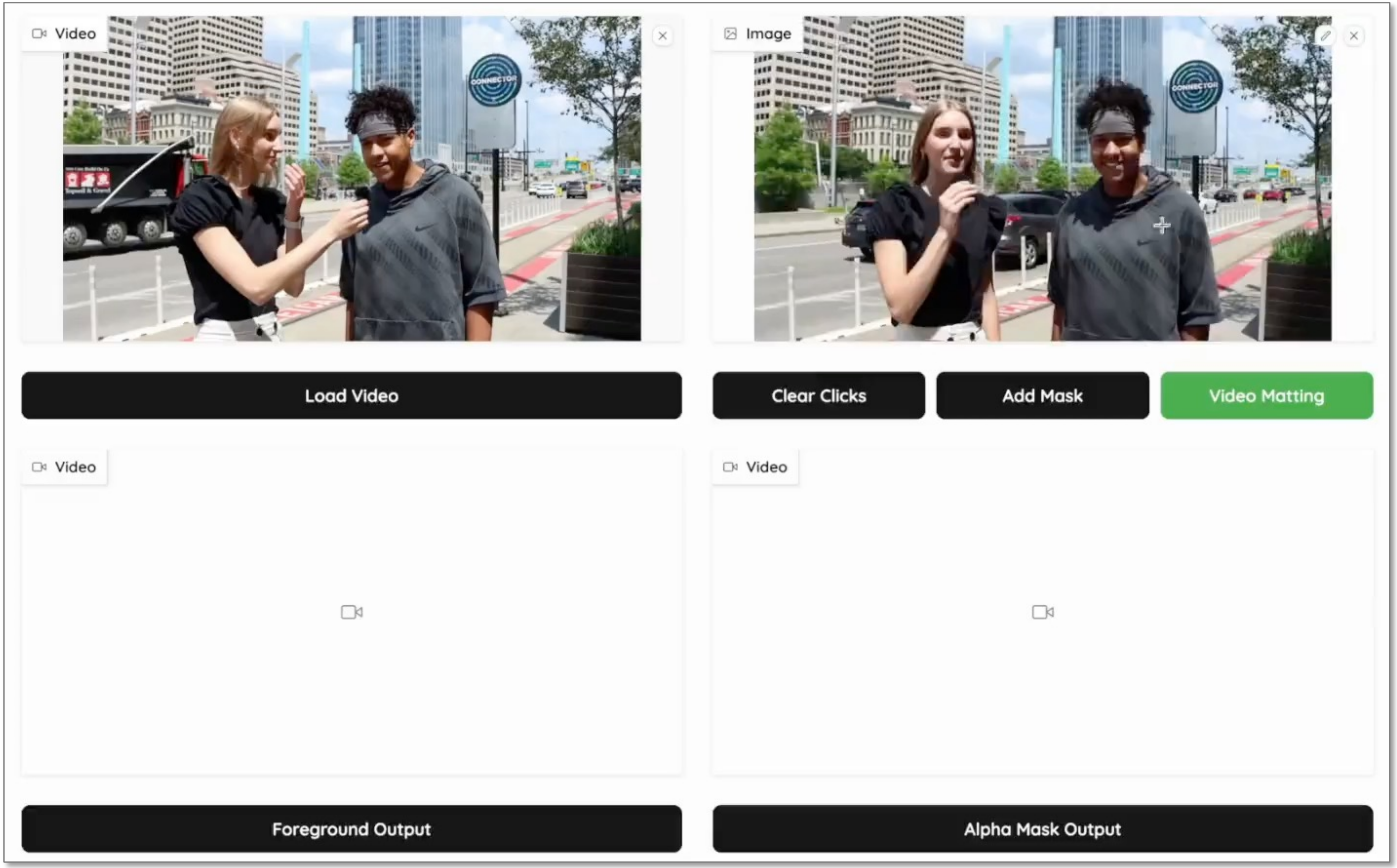
Summary

- Stable performance in both:
  - *Semantics* of core regions
  - *Fine-grained* boundary details

- Practical human video matting framework that:
  - Support target assignment
  - Increase user interactions to improve user experience

🏆 We are among the first video matting projects that provide interactive online demo that could be easily used with a few clicks.


Video
Image
Load Video
Clear Clicks
Add Mask
Video Matting
Video
Video
Foreground Output
Alpha Mask Output


Video
Image
Load Video
Clear Clicks
Add Mask
Video Matting
Video
Video
Foreground Output
Alpha Mask Output

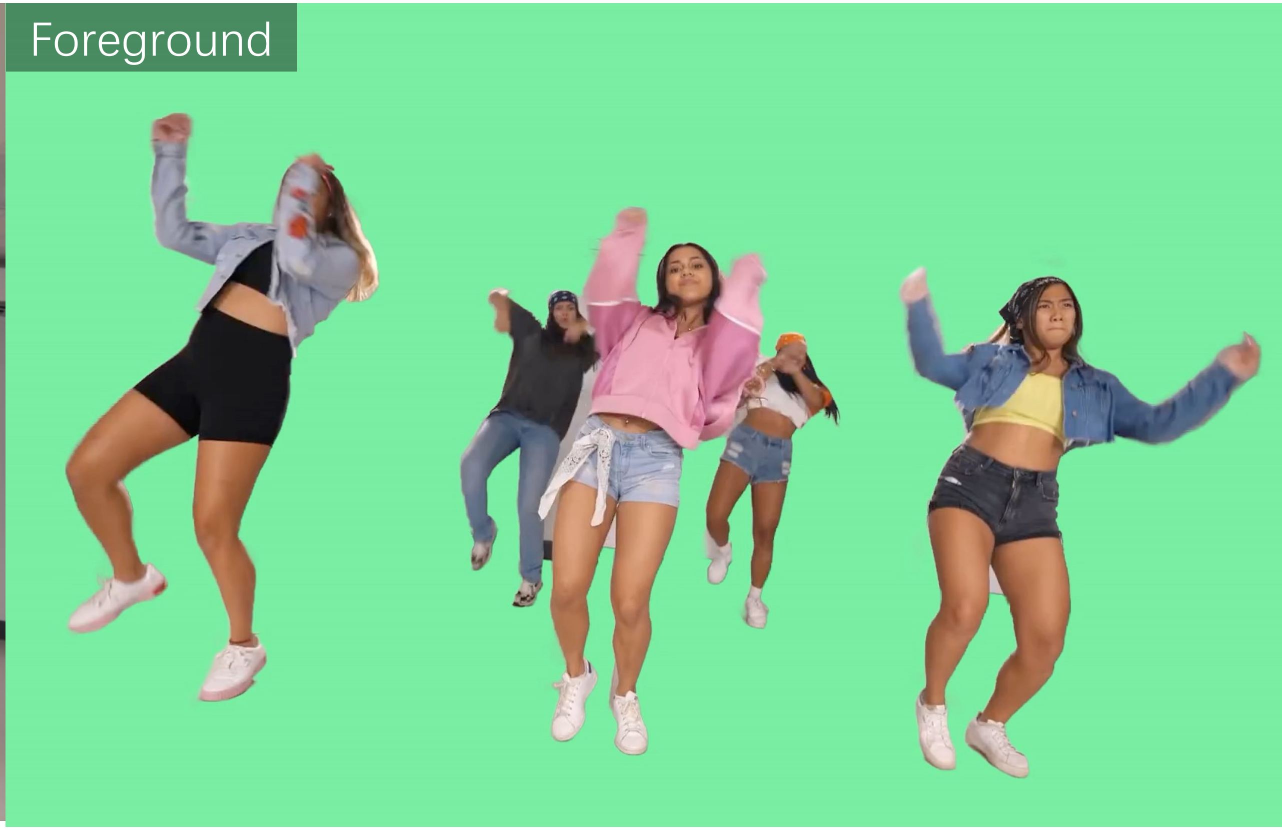
More Results on Video Matting

Videos in the Wild


Input Video

Foreground


Foreground
Input Video
Alpha Mask


Input Video

Foreground


Input Video
Foreground
Alpha Mask

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

S-LAB FOR ADVANCED INTELLIGENCE

Q&A

CVPR Nashville JUNE 11-15, 2025

MatAnyone

Stable Video Matting with Consistent Memory Propagation

Peiqing Yang[1], Shangchen Zhou[1], Jixin Zhao[1], Qingyi Tao[2], Chen Change Loy[1]

[1]S-Lab, Nanyang Technological University, [2]SenseTime Research, Singapore

CVPR 2025 🔥 **1K GitHub Stars**

Code

Demo