MaGGIe: Masked Guided Gradual Human Instance Matting

Supplementary Material

7. Architecture details

This section delves into the architectural nuances of our framework, providing a more detailed exposition of components briefly mentioned in the main paper. These insights are crucial for a comprehensive understanding of the underlying mechanisms of our approach.

7.1. Mask guidance identity embedding

We embed mask guidance into a learnable space before inputting it into our network. This approach, inspired by the ID assignment in AOT [55], generates a guidance embedding $\mathbf{E} \in \mathbb{R}^{T \times C_e \times H \times W}$ by mapping embedding vectors $\mathbf{D} \in \mathbb{R}^{N \times C_e}$ to pixels based on the guidance mask $\mathbf{M}$:

$$\mathbf{E}(x, y) = \mathbf{M}(x, y)\mathbf{D}. \tag{4}$$

Here, $\mathbf{E}(x, y) \in \mathbb{R}^{T \times C_e}$ and $\mathbf{M}(x, y) \in \{0, 1\}^{T \times N}$ represent the values at row y and column x in $\mathbf{E}$ and $\mathbf{M}$, respectively. In our experiment, we set $N = 10$, but it can be any larger number without affecting the architecture significantly.

7.2. Feature extractor

In our experiments, we employ ResNet-29 [13] as the feature extractor, consistent with other baselines [49, 56]. We have $C_8 = 128, C_4 = 64, C_1 = C_2 = 32$.

7.3. Dense-image to sparse-instance features

We express the Eq. (2) as the visualization in Fig. 7. It involves extracting feature vectors $\bar{\mathbf{F}}(x, y, t)$ and instance token vectors $\mathbf{T}_i$ for each uncertainty index $(x, y, t, i) \in \mathbf{U}$. These vectors undergo channel-wise multiplication, emphasizing channels relevant to each instance. A subsequent MLP layer then converts this product into sparse, instance-specific features.

7.4. Detail aggregation

This process, akin to a U-net decoder, aggregates features from different scales, as detailed in Fig. 8. It involves upscaling sparse features and merging them with corresponding higher-scale features. However, this requires precomputed downscale indices from dummy sparse convolutions on the full input image.

7.5. Sparse matte head

Our matte head design, inspired by MGM [56], comprises two sparse convolutions with intermediate normalization and activation (Leaky ReLU) layers. The final output undergoes sigmoid activation for the final prediction. Non-refined locations in the dense prediction are assigned a value of zero.

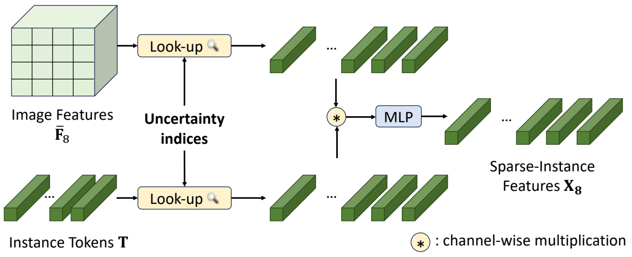
Look-up
Image Features $\bar{\mathbf{F}}_8$
Uncertainty indices
*
MLP
Sparse-Instance Features $\mathbf{X}_8$
Look-up
Instance Tokens $\mathbf{T}$
* : channel-wise multiplication

Figure 7. **Converting Dense-Image to Sparse-Instance Features**. We transform the dense image features into sparse, instance-specific features with the help of instance tokens.

7.6. Sparse progressive refinement

The PRM module progressively refines $\mathbf{A}_8 \rightarrow \mathbf{A}_4 \rightarrow \mathbf{A}_1$ to have $\mathbf{A}$. We assume that all predictions are rescaled to the largest size and perform refinement between intermediate

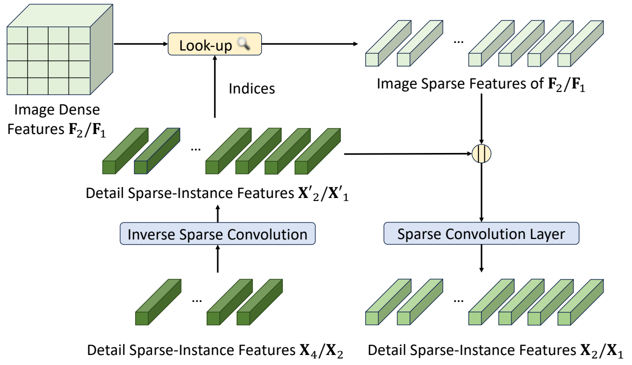
Look-up
Indices
Image Sparse Features of $\mathbf{F}_2/\mathbf{F}_1$
Image Dense Features $\mathbf{F}_2/\mathbf{F}_1$
Detail Sparse-Instance Features $\mathbf{X'}_2/\mathbf{X'}_1$
Inverse Sparse Convolution
Sparse Convolution Layer
Detail Sparse-Instance Features $\mathbf{X}_4/\mathbf{X}_2$
Detail Sparse-Instance Features $\mathbf{X}_2/\mathbf{X}_1$

Figure 8. **Detail Aggregation Module merges sparse features across scales**. This module equalizes spatial scales of sparse features using inverse sparse convolution, facilitating their combination.

predictions and uncertainty indices $\mathbf{U}$:

$$\mathbf{A} = \mathbf{A}_8 \tag{5}$$

$$\mathbf{R}_4(j) = \begin{cases} 1, \text{ if } j \in \mathcal{D}(\mathbf{A}) \text{ and } j \in \mathbf{U} \\ 0, \text{ otherwise} \end{cases} \tag{6}$$

$$\mathbf{A} = \mathbf{A} \times (1 - \mathbf{R}_4) + \mathbf{R}_4 \times \mathbf{A}_4 \tag{7}$$

$$\mathbf{R}_1(j) = \begin{cases} 1, \text{ if } j \in \mathcal{D}(\mathbf{A}) \text{ and } j \in \mathbf{U} \\ 0, \text{ otherwise} \end{cases} \tag{8}$$

$$\mathbf{A} = \mathbf{A} \times (1 - \mathbf{R}_1) + \mathbf{R}_1 \times \mathbf{A}_4 \tag{9}$$

where $j = (x, y, t, i)$ is an index in the output; $\mathbf{R}_1, \mathbf{R}_4$ in shape $T \times N \times H \times W$; and $\mathcal{D}(\mathbf{A}) = \text{dilation}(0 < \mathbf{A} < 1)$ is the indices of all dilated uncertainty values on $\mathbf{A}$. The dilation kernel is set to 30, 15 for $\mathbf{R}_4, \mathbf{R}_1$ respectively.

7.7. Attention loss and loss weight

With $\mathbf{A}^{gt}$ as the ground-truth alpha matte and its $\frac{1}{8}$ downscaled version $\mathbf{A}_8^{gt}$, we define a binarized $\tilde{\mathbf{A}}_8^{gt} = \mathbf{A}_8^{gt} > 0$. The attention loss $\mathcal{L}_{\text{att}}$ is:

$$\mathcal{L}_{\text{att}} = \sum_i^N \left\| \mathbf{1} - \text{Aff}(i)^\top \tilde{\mathbf{A}}_8^{gt}(i) \right\|_1 \tag{10}$$

aiming to maximize each instance token $\mathbf{T}_i$'s attention score to its corresponding groundtruth region $\tilde{\mathbf{A}}_8^{gt}(i)$.

The weight $\mathbf{W}_8$ at each location is:

$$\mathbf{W}_8(j) = \begin{cases} \gamma, \text{ if } 0 < \mathbf{A}_8^{gt}(j) < 1 \text{ and } 0 < \mathbf{A}_8(j) < 1 \\ 1.0, \text{ otherwise} \end{cases} \tag{11}$$

with $\gamma = 2.0$ in our experiments, focusing on under-refined ground-truth and over-refined predicted areas.

7.8. Temporal sparsity prediction

A key aspect of our approach is the prediction of temporal sparsity to maintain consistency between frames. This module contrasts the feature maps of consecutive frames to predict their absolute differences. Comprising three convolution layers with batch normalization and ReLU activation, this module processes the concatenated feature maps from two adjacent frames and predicts the binary differences between them.

Unlike SparseMat [50], which relies on manual threshold selection for frame differences, our method offers a more robust and domain-independent approach to determining frame sparsity. This is particularly effective in handling variations in movement, resolution, and domain between frames, as demonstrated in Fig. 9

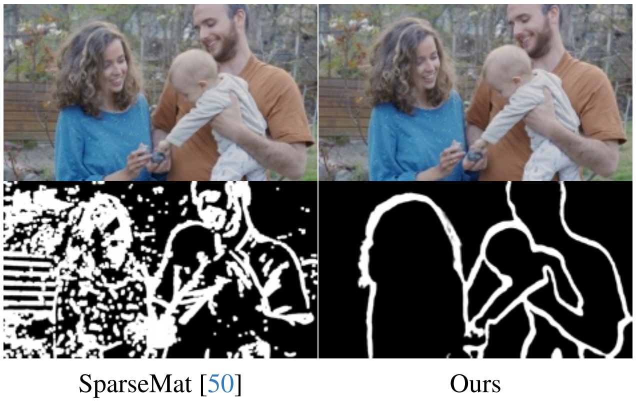

Figure 9. **Temporal Sparsity Between Two Consecutive Frames**. The top row displays a pair of successive frames. Below, the second row illustrates the predicted differences by two distinct frameworks, with areas of discrepancy emphasized in white. In contrast to SparseMat's output, which appears cluttered and noisy, our module generates a more refined sparsity map. This map effectively accentuates the foreground regions that undergo notable changes between the frames, providing a clearer and more focused representation of temporal sparsity. (Best viewed in color).

7.9. Forward and backward matte fusion

The forward-backward fusion for the i-th instance at frame t is respectively:

$$\begin{aligned} \mathbf{A}^f(t, i) &= \Delta(t) \times \mathbf{A}(t, i) \\ &\quad + (1 - \Delta(t)) \times \mathbf{A}^f(t-1, i) \end{aligned} \tag{12}$$

$$\begin{aligned} \mathbf{A}^b(t, i) &= \Delta(t+1) \times \mathbf{A}(t, i) \\ &\quad + (1 - \Delta(t+1)) \times \mathbf{A}^b(t+1, i) \end{aligned} \tag{13}$$

Each entry $j = (x, y, t, i)$ on final output $\mathbf{A}^{\text{temp}}$ is:

$$\mathbf{A}^{\text{temp}}(j) = \begin{cases} \mathbf{A}(j), \text{ if } \mathbf{A}^f(j) \neq \mathbf{A}^b(j) \\ \mathbf{A}^f(j), \text{ otherwise} \end{cases} \tag{14}$$

This fusion enhances temporal consistency and minimizes error propagation.

8. Image matting

This section expands on the image matting process, providing additional insights into dataset generation and comprehensive comparisons with existing methods. We delve into the creation of I-HIM50K and M-HIM2K datasets, offer detailed quantitative analyses, and present further qualitative results to underscore the effectiveness of our approach.

8.1. Dataset generation and preparation

The I-HIM50K dataset was synthesized from the HHM50K [50] dataset, which is known for its extensive collection of human image mattes. We employed a

Table 8. **Ten models with vary mask quality are used in M-HIM2K.** The MaskRCNN models are from detectron2 trained on COCO with different settings.

Model	COCO mask AP (%)
r50_c4_3x	34.4
r50_dc5_3x	35.9
r101_c4_3x	36.7
r50_fpn_3x	37.2
r101_fpn_3x	38.6
x101_fpn_3x	39.5
r50_fpn_400e	42.5
regnety_400e	43.3
regnetx_400e	43.5
r101_fpn_400e	43.7

MaskRCNN [14] Resnet-50 FPN 3x model, trained on the COCO dataset, to filter out single-person images, resulting in a subset of 35,053 images. Following the InstMatt [49] methodology, these images were composited against diverse backgrounds from the BG20K [29] dataset, creating multi-instance scenarios with 2-5 subjects per image. The subjects were resized and positioned to maintain a realistic scale and avoid excessive overlap, as indicated by instance IoUs not exceeding 30%. This process yielded 49,737 images, averaging 2.28 instances per image. During training, guidance masks were generated by binarizing the alpha mattes and applying random dropout, dilation, and erosion operations. Sample images from I-HIM50K are displayed in Fig. 10.

The M-HIM2K dataset was designed to test model robustness against varying mask qualities. It comprises ten masks per instance, generated using various MaskRCNN models. More information about models used for this generation process is shown in Table 8. The masks were matched to instances based on the highest IoU with the ground truth alpha mattes, ensuring a minimum IoU threshold of 70%. Masks that did not meet this threshold were artificially generated from ground truth. This process resulted in a comprehensive set of 134,240 masks, with 117,660 for composite and 16,600 for natural images, providing a robust benchmark for evaluating masked guided instance matting. The full dataset I-HIM50K and M-HIM2K will be released after the acceptance of this work.

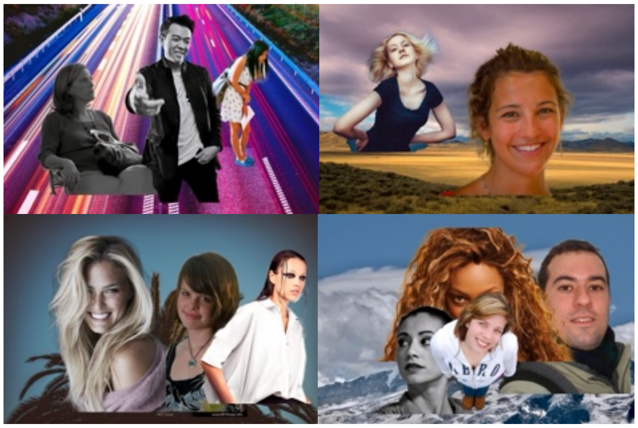

Figure 10. **Examples of I-HIM50K dataset**. (Best viewed in color).

8.2. Training details

We initialized our feature extractor with ImageNet [43] weights, following previous methods [49, 56]. Our models were retrained on the I-HIM50K dataset with a crop size 512. All baselines underwent 100 training epochs, using the HIM2K composition set for validation. The training was conducted on 4 A100 GPUs with a batch size 96. We employed AdamW for optimization, with a learning rate of 1.5×10^{-4} and a cosine decay schedule post 1,500 warm-up iterations. The training also incorporated curriculum learning as MGM and standard augmentation as other baselines. During training, mask orders were shuffled, and some masks were randomly omitted. In testing, images were resized to have a short side of 576 pixels.

8.3. Quantitative details

We extend the ablation study from the main paper, providing detailed statistics in Table 9 and Table 10. These tables offer insights into the average and standard deviation

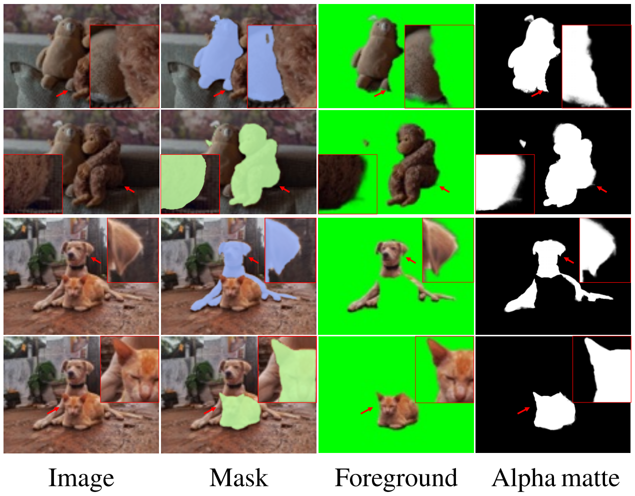

Figure 11. **Our framework can generalize to any object.** Without humans appearing in the image, our framework still performs the matting task very well to the mask-guided objects. (Best viewed in color and digital zoom).

Table 9. **Full details of different input mask setting on HIM2K+M-HIM2K.** (Extension of Table 3). **Bold** denotes the lowest average error.

Mask input	Composition							Natural							
	MAD	MAD$_f$	MAD$_u$	MSE	SAD	Grad	Conn	MAD	MAD$_f$	MAD$_u$	MSE	SAD	Grad	Conn	
Stacked	27.01	68.83	381.27	18.82	16.35	16.80	15.72	39.29	61.39	213.27	25.10	25.52	16.44	23.26	mean
	0.83	5.93	7.06	0.76	0.50	0.31	0.51	4.21	13.37	14.10	4.01	2.00	0.70	2.02	std
Embeded($C_e = 1$)	19.18	68.04	330.06	12.40	11.64	13.00	11.16	33.60	60.35	**188.44**	20.63	21.40	13.44	19.18	mean
	0.87	8.07	6.96	0.80	0.52	0.27	0.52	4.07	12.60	12.28	3.86	1.81	0.57	1.83	std
Embeded($C_e = 2$)	21.74	84.64	355.95	14.46	13.23	14.39	12.69	35.16	59.55	193.95	21.93	22.59	14.51	20.40	mean
	0.92	7.33	7.68	0.85	0.55	0.27	0.55	4.23	13.79	13.45	4.03	2.31	0.61	2.32	std
Embeded($C_e = 3$)	**17.75**	**53.23**	**315.43**	**11.19**	**10.79**	**12.52**	**10.32**	**33.06**	**56.69**	189.59	**20.22**	**19.43**	**13.11**	**17.30**	mean
	0.66	5.04	6.31	0.60	0.39	0.24	0.39	3.74	11.90	12.49	3.58	1.92	0.51	1.95	std
Embeded($C_e = 5$)	24.79	73.22	384.14	17.07	15.09	16.19	14.58	34.25	65.57	216.56	20.39	21.89	15.66	19.70	mean
	0.88	4.99	7.24	0.79	0.52	0.30	0.52	4.16	13.59	13.09	3.96	2.31	0.58	2.32	std

Table 10. **Full details of different training objective components on HIM2K+M-HIM2K.** (Extension of Table 4). **Bold** denotes the lowest average error.

$\mathcal{L}_{att}$	$\mathbf{W}_8$	Composition							Natural							
		MAD	MAD$_f$	MAD$_u$	MSE	SAD	Grad	Conn	MAD	MAD$_f$	MAD$_u$	MSE	SAD	Grad	Conn	
		31.77	52.70	**294.22**	24.13	18.92	16.58	18.27	46.68	51.23	**176.60**	33.61	32.89	15.68	30.64	mean
		0.90	4.92	5.24	0.85	0.54	0.26	0.54	3.64	10.27	9.58	3.47	1.85	0.50	1.85	std
	✓	25.41	104.24	342.19	18.36	15.29	14.53	14.75	46.30	87.18	210.72	32.93	31.40	15.84	29.26	mean
		0.72	6.15	5.53	0.67	0.43	0.23	0.43	3.71	11.68	10.62	3.55	1.85	0.50	1.86	std
✓		17.56	53.51	302.07	11.24	**10.65**	**12.34**	10.22	32.95	**51.11**	183.13	20.41	**19.23**	13.29	**17.06**	mean
		0.75	6.32	6.32	0.70	0.45	0.27	0.45	3.34	10.25	10.99	3.19	2.04	0.60	2.06	std
✓	✓	**17.55**	**47.81**	301.96	**11.23**	10.68	**12.34**	**10.19**	**32.03**	53.15	183.42	**19.42**	19.60	**13.16**	17.43	mean
		0.68	5.21	5.73	0.63	0.41	0.25	0.41	3.48	10.77	11.18	3.32	1.92	0.55	1.94	std

of performance metrics across HIM2K [49] and M-HIM2K datasets. Our model not only achieves competitive average results but also maintains low variability in performance across different error metrics. Additionally, we include the Sum Absolute Difference (SAD) metric, aligning with previous image matting benchmarks.

Comprehensive quantitative results comparing our model with baseline methods on HIM2K and M-HIM2K are presented in Table 12. This analysis highlights the impact of mask quality on matting output, with our model demonstrating consistent performance even with varying mask inputs.

We also perform another experiment when the MGM-style refinement replaces our proposed sparse guided progressive refinement. The Table 11 shows the results where our proposed method outperforms the previous approach in all metrics.

8.4. More qualitative results on natural images

Fig. 13 showcases our model's performance in challenging scenarios, particularly in accurately rendering hair regions. Our framework consistently outperforms MGM* in detail preservation, especially in complex instance interactions. In comparison with InstMatt, our model exhibits superior instance separation and detail accuracy in ambiguous regions.

Fig. 14 and Fig. 15 illustrate the performance of our model and previous works in extreme cases involving multiple instances. While MGM* struggles with noise and accuracy in dense instance scenarios, our model maintains high precision. InstMatt, without additional training data, shows limitations in these complex settings.

Table 11. **Compare between previous dense progressive refinement (PR) - MGM and our proposed guided sparse progressive refinement.** Numbers are mean on HIM2K+M-HIM2K and small numbers indicate the std.

PR	MAD	MSE	Grad	Conn	MAD$_f$	MAD$_u$
Comp Set						
MGM	14.70 (0.4)	8.87 (0.3)	10.39 (0.2)	8.44 (0.2)	32.02 (1.5)	252.34 (4.2)
Ours	**12.93 (0.3)**	**7.26 (0.3)**	**8.91 (0.1)**	**7.37 (0.2)**	**19.54 (1.0)**	**235.95 (3.4)**
Natural Set						
MGM	27.66 (4.1)	16.94 (3.9)	10.49 (0.7)	13.95 (1.5)	52.72 (12.1)	150.71 (13.3)
Ours	**27.17 (3.3)**	**16.09 (3.2)**	**9.94 (0.6)**	**13.42 (1.4)**	**49.52 (8.0)**	**146.71 (11.6)**

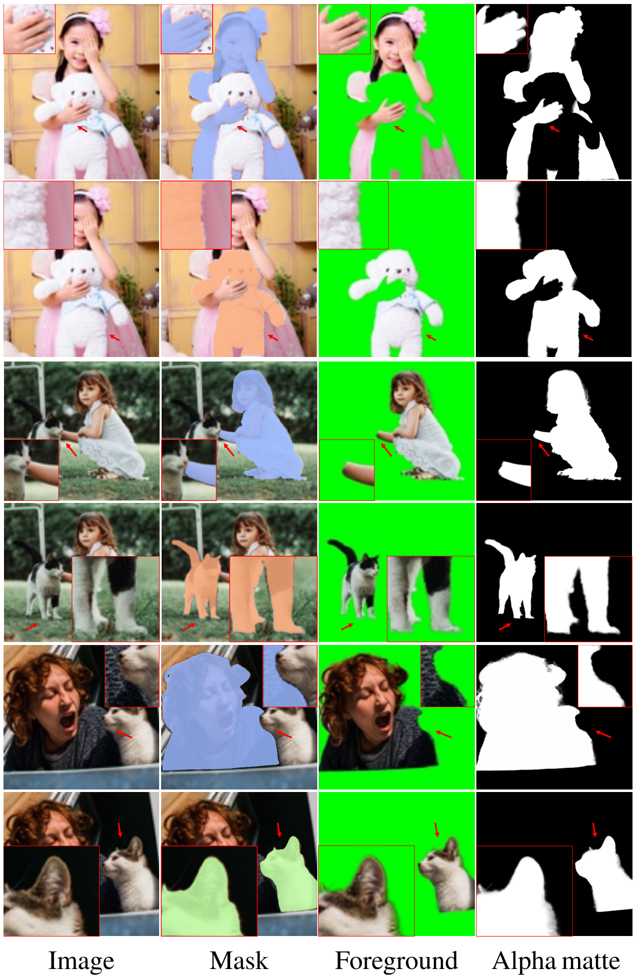

Figure 12. **Our solution is not limited to human instances.** When testing with other objects, our solution is able to produce fairly accurate alpha matte without training on them. (Best viewed in color and digital zoom).

The robustness of our mask-guided approach is further demonstrated in Fig. 16. Here, we highlight the challenges faced by MGM variants and SparseMat in predicting missing parts in mask inputs, which our model addresses. However, it is important to note that our model is not designed as a human instance segmentation network. As shown in Fig. 17, our framework adheres to the input guidance, ensuring precise alpha matte prediction even with multiple instances in the same mask.

Lastly, Fig. 12 and Fig. 11 emphasize our model's generalization capabilities. The model accurately extracts both human subjects and other objects from backgrounds, showcasing its versatility across various scenarios and object types.

All examples are Internet images without groundtruth and the mask from r101_fpn_400e are used as the guidance.

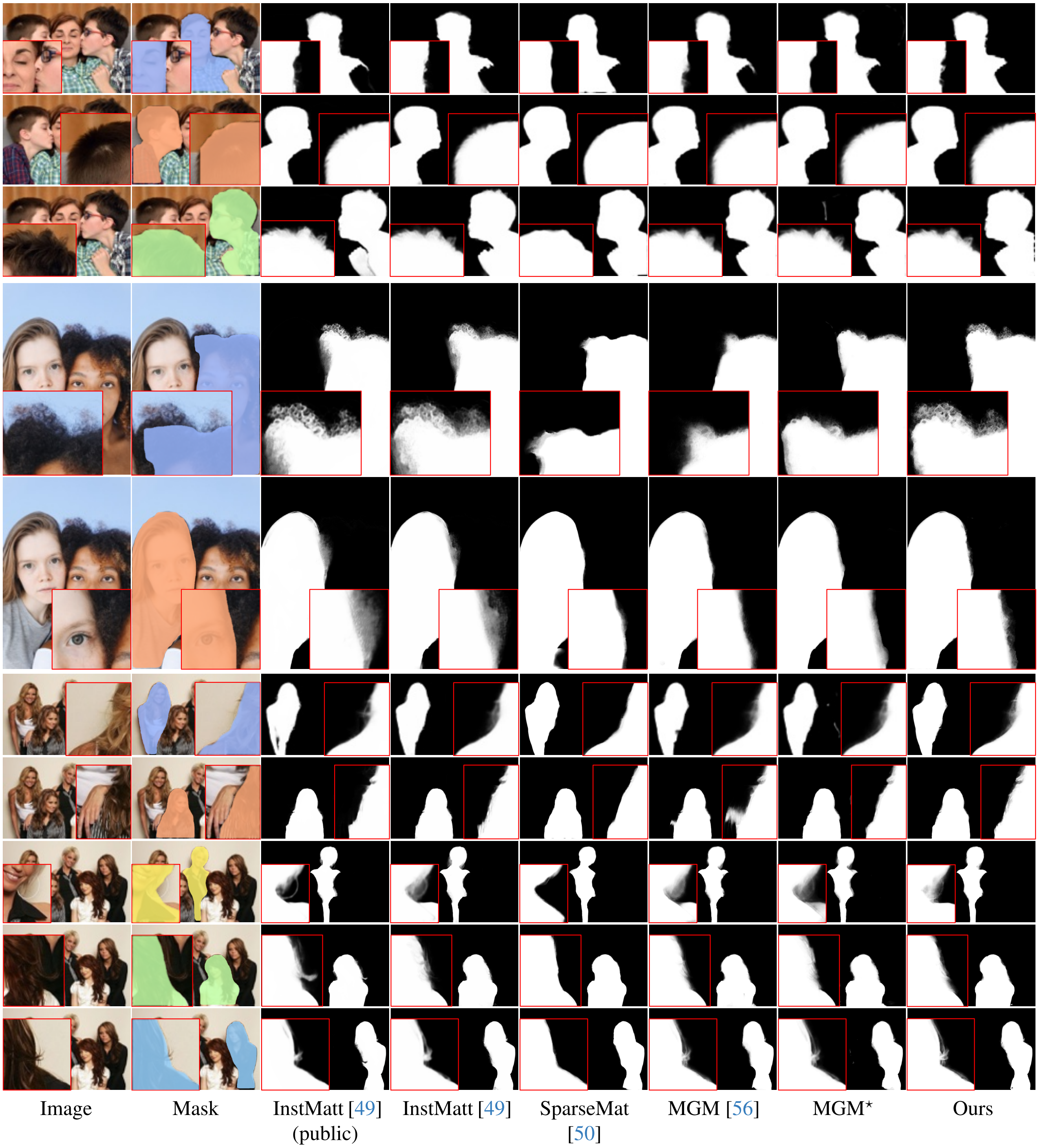

Figure 13. **Our model produces highly detailed alpha matte on natural images.** Our results show that it is accurate and comparable with previous instance-agnostic and instance-awareness methods without expensive computational costs. Red squares zoom in the detail regions for each instance. (Best viewed in color and digital zoom).

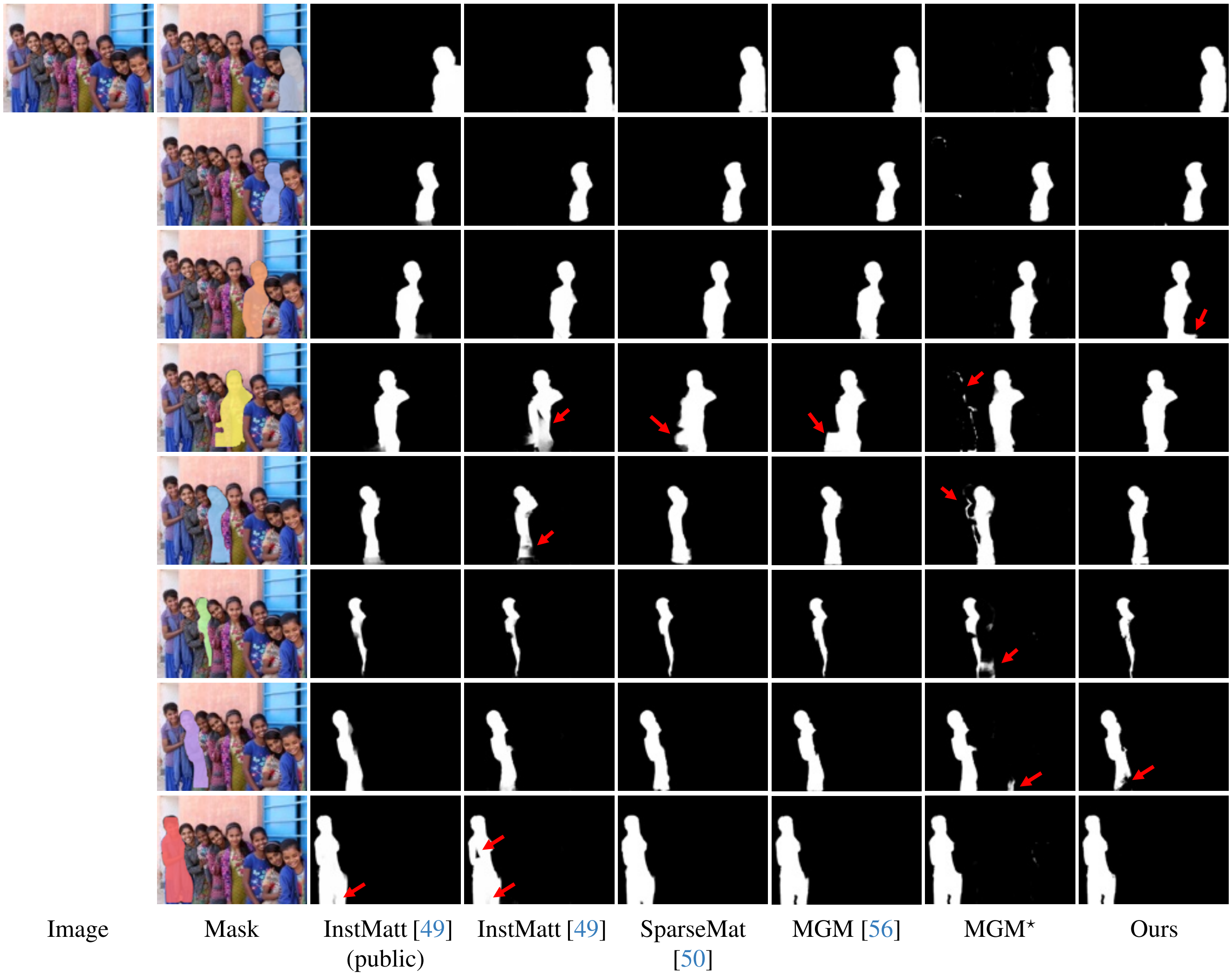

Figure 14. **Our frameworks precisely separate instances in an extreme case with many instances.** While MGM often causes the overlapping between instances and MGM* contains noises, ours produces on-par results with InstMatt trained on the external dataset. Red arrow indicates the errors. (Best viewed in color and digital zoom).

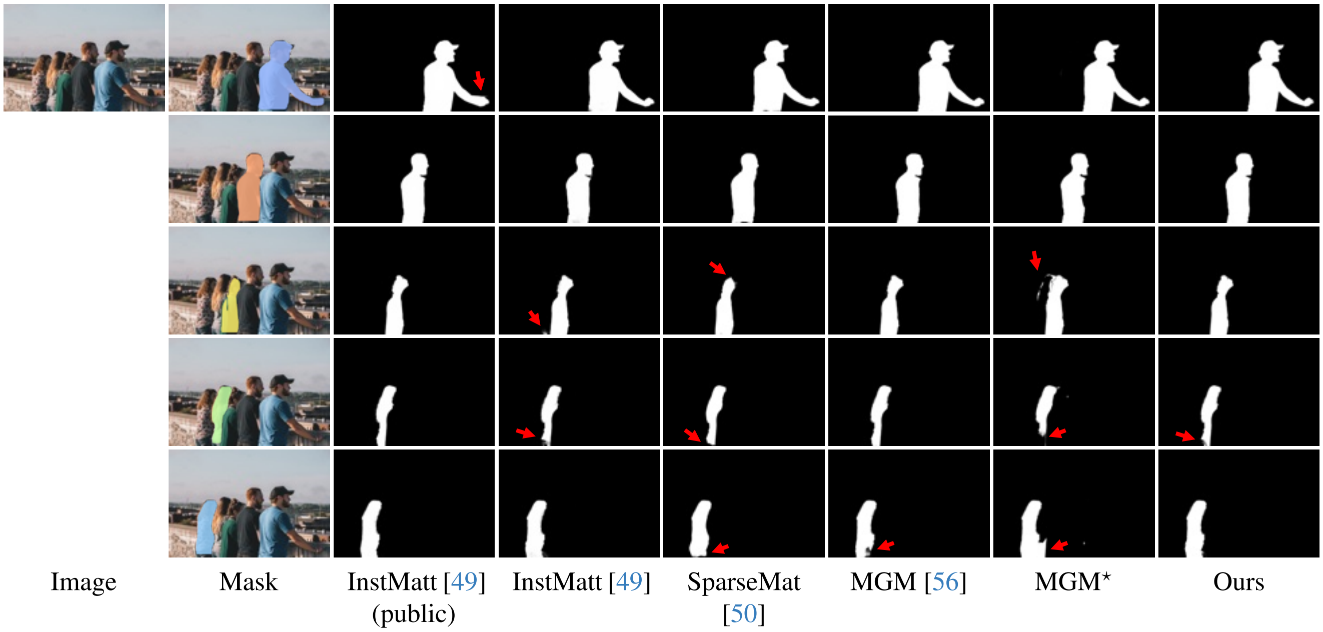

Figure 15. **Our frameworks precisely separate instances in a single pass.** The proposed solution shows comparable results with InstMatt and MGM without running the prediction/refinement five times. Red arrow indicates the errors. (Best viewed in color and digital zoom).

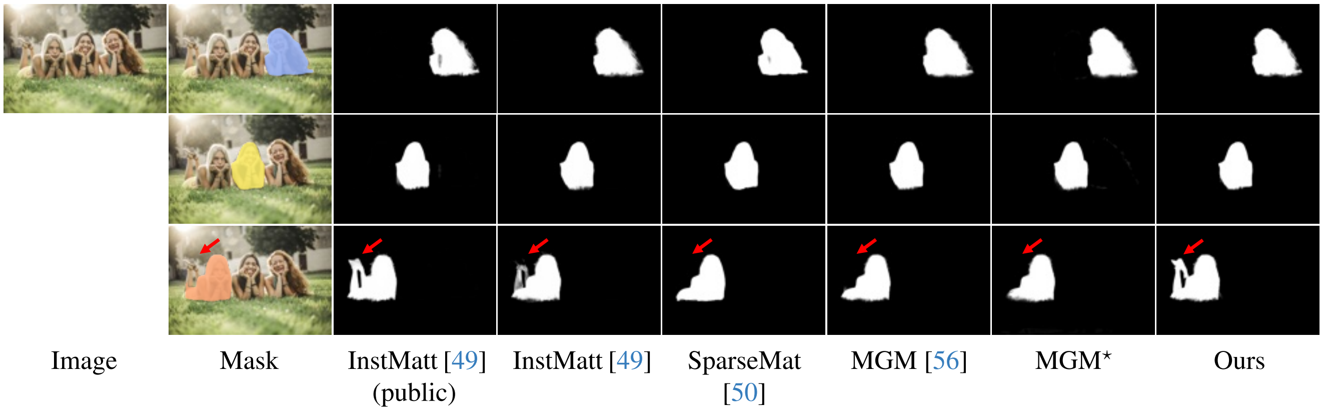

Figure 16. **Unlike MGM and SparseMat, our model is robust to the input guidance mask.** With the attention head, our model produces more stable results to mask inputs without complex refinement between instances like InstMatt. Red arrow indicates the errors. (Best viewed in color and digital zoom).

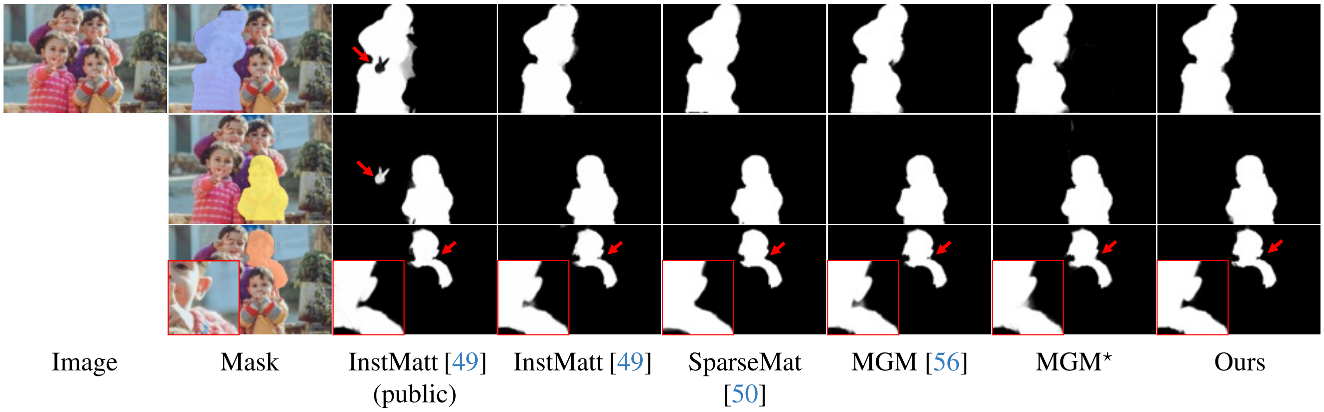

Figure 17. **Our solution works correctly with multi-instance mask guidances.** When multiple instances exist in one guidance mask, we still produce the correct union alpha matte for those instances. Red arrow indicates the errors or the zoom-in region in red box. (Best viewed in color and digital zoom).

Table 12. **Details of quantitative results on HIM2K+M-HIM2K** (Extension of Table 5). Gray indicates the public weight without retraining.

Model	Composition set							Natural set							Mask from
	MAD	MAD_f	MAD_u	MSE	SAD	Grad	Conn	MAD	MAD_f	MAD_u	MSE	SAD	Grad	Conn	
Instance-agnostic															
MGM [39]	25.79	69.67	331.73	17.00	15.65	13.64	14.91	48.05	103.81	233.85	32.66	27.44	14.72	25.07	r50_c4_3x
	24.75	70.92	316.59	16.21	15.01	13.17	14.23	34.67	66.28	183.48	21.03	22.82	12.79	20.30	r50_dc5_3x
	23.60	66.79	321.23	15.03	14.38	13.19	13.62	35.51	70.94	198.99	20.96	22.62	13.73	20.17	r101_c4_3x
	24.55	67.27	316.29	15.97	14.91	13.14	14.12	33.66	67.41	184.99	19.93	21.99	13.06	19.43	r50_fpn_3x
	23.42	66.37	310.99	14.94	14.21	12.84	13.42	35.14	72.30	183.87	21.02	21.87	12.82	19.34	r101_fpn_3x
	22.71	63.35	305.67	14.36	13.81	12.64	13.03	31.06	61.76	175.33	17.60	20.98	12.61	18.44	x101_fpn_3x
	22.03	61.91	300.29	13.85	13.36	12.30	12.59	29.16	57.59	165.22	15.93	20.10	11.76	17.56	r50_fpn_400e
	21.37	57.28	296.73	13.18	12.98	12.16	12.21	26.40	51.24	158.95	13.42	17.73	11.45	15.10	regnety_400e
	21.78	60.31	297.14	13.62	13.22	12.25	12.46	27.09	49.26	160.05	13.82	17.48	11.20	14.87	regnetx_400e
	21.52	60.07	297.14	13.44	13.14	12.20	12.38	24.41	51.46	152.90	11.62	17.43	11.09	14.84	r101_fpn_400e
	23.15	64.39	309.38	14.76	14.07	12.75	13.30	32.52	65.20	179.76	18.80	21.05	12.52	18.51	**mean**
	1.52	4.49	12.01	1.30	0.92	0.52	0.92	6.74	15.94	23.87	5.99	3.09	1.17	3.16	**std**
MGM [56]	15.94	32.55	266.64	9.62	9.68	10.11	9.18	37.55	86.64	191.09	24.03	21.15	11.34	18.94	r50_c4_3x
	16.05	36.36	264.96	9.81	9.75	10.10	9.26	32.58	68.52	172.83	19.58	20.17	10.92	17.80	r50_dc5_3x
	15.40	30.89	264.28	9.17	9.37	10.01	8.90	31.24	69.59	175.67	18.15	18.57	10.83	16.26	r101_c4_3x
	15.93	34.54	265.44	9.68	9.67	10.10	9.20	32.83	75.06	173.63	19.72	19.13	10.85	16.81	r50_fpn_3x
	15.74	34.23	263.35	9.50	9.55	10.02	9.07	30.77	69.10	171.92	17.78	18.22	10.67	15.95	r101_fpn_3x
	15.23	36.18	260.80	9.03	9.27	9.92	8.76	30.09	63.23	167.58	17.34	18.51	10.69	16.09	x101_fpn_3x
	14.96	34.13	259.17	8.81	9.08	9.83	8.61	28.28	50.35	158.02	15.71	17.71	10.24	15.25	r50_fpn_400e
	14.53	31.71	256.33	8.41	8.83	9.73	8.35	26.95	49.55	155.63	14.43	15.69	9.98	13.34	regnety_400e
	14.82	33.06	257.09	8.69	9.01	9.80	8.53	26.61	47.81	154.05	14.22	15.45	9.87	13.16	regnetx_400e
	14.65	31.71	256.29	8.53	8.94	9.74	8.46	25.42	51.73	153.11	13.03	15.73	9.90	13.44	r101_fpn_400e
	15.32	33.54	261.43	9.13	9.31	9.94	8.83	30.23	63.16	167.35	17.40	18.03	10.53	15.70	**mean**
	0.57	1.88	4.00	0.51	0.34	0.15	0.34	3.62	12.97	12.14	3.26	1.93	0.50	1.94	**std**

Table 12. **Details of quantitative results on HIM2K+M-HIM2K** (Extension of Table 5). Gray indicates the public weight without retraining. (Continued)

SparseMat [50]	23.14	47.59	378.89	16.37	13.97	15.56	13.54	46.28	101.48	255.98	31.99	26.81	17.97	24.82	r50_c4_3x
	21.94	49.48	358.08	15.36	13.24	14.90	12.80	36.93	67.62	213.46	23.76	22.11	16.05	20.01	r50_dc5_3x
	21.78	43.36	368.59	15.15	13.16	15.21	12.72	38.32	77.98	234.69	24.51	22.83	17.19	20.78	r101_c4_3x
	21.94	47.00	361.30	15.33	13.24	14.99	12.80	37.16	74.18	218.62	23.95	21.95	16.39	19.86	r50_fpn_3x
	21.43	46.51	356.43	14.88	12.93	14.81	12.48	35.95	72.78	218.46	22.62	20.67	16.11	18.58	r101_fpn_3x
	20.63	47.73	349.81	14.12	12.48	14.58	12.02	34.32	64.51	209.64	21.10	20.44	16.03	18.33	x101_fpn_3x
	20.29	44.20	342.14	13.93	12.22	14.21	11.76	31.44	57.51	197.53	18.58	19.49	14.96	17.35	r50_fpn_400e
	19.65	41.20	340.38	13.29	11.85	14.08	11.38	30.21	48.53	194.90	17.32	17.47	14.82	15.31	regnety_400e
	19.90	41.40	336.40	13.56	12.02	14.03	11.56	29.85	52.17	191.09	16.99	17.19	14.52	15.03	regnetx_400e
	19.81	43.43	337.43	13.50	12.01	14.05	11.55	29.83	61.40	191.89	17.07	17.13	14.48	14.96	r101_fpn_400e
	21.05	45.19	352.95	14.55	12.71	14.64	12.26	35.03	67.82	212.63	21.79	20.61	15.85	18.50	**mean**
	1.17	2.85	14.24	1.02	0.70	0.54	0.71	5.13	15.19	20.77	4.68	3.03	1.16	3.08	**std**
Instance-awareness															
InstMatt [49]	12.98	23.71	257.74	5.76	7.94	9.47	7.27	31.15	60.03	174.10	15.91	18.12	10.64	15.73	r50_c4_3x
	13.15	23.08	257.38	5.96	8.05	9.48	7.38	28.05	51.53	164.19	13.63	16.89	10.33	14.53	r50_dc5_3x
	12.99	22.42	257.52	5.79	7.93	9.47	7.26	27.06	48.52	162.72	12.90	16.06	10.29	13.68	r101_c4_3x
	13.13	20.60	256.70	5.90	8.03	9.47	7.36	28.31	49.87	164.16	13.97	16.86	10.37	14.49	r50_fpn_3x
	13.04	23.98	257.51	5.85	7.96	9.45	7.28	28.92	59.32	168.72	14.37	16.98	10.40	14.64	r101_fpn_3x
	12.77	22.16	255.33	5.63	7.83	9.40	7.16	27.02	46.39	162.89	12.82	16.49	10.27	14.08	x101_fpn_3x
	12.61	21.31	254.27	5.55	7.71	9.36	7.05	25.33	44.84	157.03	11.23	15.54	9.97	13.18	r50_fpn_400e
	12.58	23.53	253.85	5.57	7.69	9.35	7.03	24.34	41.62	154.89	10.65	15.22	10.00	12.85	regnety_400e
	12.59	20.48	252.68	5.53	7.71	9.35	7.04	24.18	40.96	154.69	10.09	14.68	9.82	12.28	regnetx_400e
	12.67	21.14	253.13	5.60	7.75	9.35	7.09	23.22	43.23	151.78	9.67	15.00	9.88	12.60	r101_fpn_400e
	12.85	22.24	255.61	5.71	7.86	9.41	7.19	26.76	48.63	161.52	12.52	16.18	10.20	13.81	**mean**
	0.23	1.31	2.00	0.16	0.14	0.06	0.13	2.48	6.76	6.94	2.05	1.08	0.26	1.08	**std**
InstMatt [49]	18.23	57.23	298.66	10.51	11.06	11.33	10.45	37.91	86.84	202.20	22.28	21.31	12.22	19.11	r50_c4_3x
	17.85	58.98	291.50	10.38	10.87	11.13	10.27	30.10	63.83	173.94	15.90	18.01	11.25	15.82	r50_dc5_3x
	17.25	51.21	292.66	9.80	10.50	11.13	9.90	30.22	59.65	178.94	15.62	17.49	11.55	15.23	r101_c4_3x
	17.69	55.80	292.90	10.22	10.80	11.19	10.19	30.27	60.16	175.66	16.44	17.38	11.33	15.13	r50_fpn_3x
	17.18	55.67	288.95	9.85	10.45	11.02	9.84	28.80	60.88	170.89	14.55	16.88	11.12	14.69	r101_fpn_3x
	16.65	53.37	284.66	9.41	10.16	10.85	9.56	27.77	55.06	168.20	14.14	16.91	11.04	14.70	x101_fpn_3x
	16.29	52.00	281.15	9.21	9.88	10.69	9.29	25.51	52.89	156.40	12.15	15.90	10.47	13.70	r50_fpn_400e
	15.99	50.92	279.15	8.97	9.71	10.65	9.12	24.82	45.83	156.46	11.83	15.14	10.43	12.94	regnety_400e
	16.47	51.85	280.00	9.37	10.01	10.69	9.42	23.73	47.85	153.70	10.35	14.69	10.17	12.49	regnetx_400e
	16.30	50.58	279.40	9.29	9.95	10.63	9.36	22.47	45.33	150.96	9.72	14.71	10.17	12.50	r101_fpn_400e
	16.99	53.76	286.90	9.70	10.34	10.93	9.74	28.16	57.83	168.74	14.30	16.84	10.98	14.63	**mean**
	0.76	2.96	6.95	0.53	0.47	0.26	0.46	4.45	12.15	15.45	3.65	1.97	0.66	1.97	**std**

Continued on next page

Table 12. **Details of quantitative results on HIM2K+M-HIM2K** (Extension of Table 5). Gray indicates the public weight without retraining. (Continued)

MGM*	14.87	46.70	256.01	8.32	8.99	10.31	8.32	37.36	65.40	181.68	23.97	20.50	11.66	17.45	r50_c4_3x
	14.65	43.00	253.75	8.21	8.87	10.25	8.22	33.70	60.48	172.03	20.83	18.51	11.29	15.93	r50_dc5_3x
	14.36	38.88	252.30	7.89	8.71	10.19	8.04	33.95	60.54	173.47	20.59	17.94	11.24	15.30	r101_c4_3x
	14.68	44.85	254.50	8.21	8.88	10.24	8.22	33.29	54.82	170.89	20.21	18.28	11.27	15.55	r50_fpn_3x
	14.70	44.68	254.29	8.24	8.89	10.21	8.25	32.07	68.47	171.41	18.80	17.44	11.07	14.84	r101_fpn_3x
	14.27	43.56	251.19	7.83	8.68	10.13	8.00	30.96	50.90	166.14	18.02	17.53	11.07	14.91	x101_fpn_3x
	13.94	38.70	248.02	7.58	8.46	10.00	7.79	29.86	48.23	158.22	16.92	16.91	10.79	14.32	r50_fpn_400e
	13.57	39.12	246.18	7.24	8.21	9.89	7.56	28.53	46.70	156.07	15.84	15.98	10.52	13.38	regnety_400e
	14.11	41.69	247.92	7.75	8.57	10.00	7.91	27.17	41.88	150.59	14.42	15.35	10.36	12.75	regnetx_400e
	13.95	38.26	246.60	7.60	8.48	9.95	7.83	26.89	41.53	150.85	14.23	15.74	10.42	13.12	r101_fpn_400e
	14.31	41.94	251.08	7.89	8.67	10.12	8.01	31.38	53.89	165.13	18.38	17.42	10.97	14.75	**mean**
	0.42	3.05	3.63	0.35	0.24	0.15	0.24	3.34	9.56	10.59	3.11	1.53	0.43	1.43	**std**
Ours	13.13	17.81	239.98	7.41	7.92	9.05	7.47	34.54	64.64	171.51	23.05	18.36	11.02	16.23	r50_c4_3x
	13.28	21.29	238.15	7.61	8.03	9.03	7.58	27.66	52.90	149.52	16.56	16.05	10.15	13.90	r50_dc5_3x
	13.20	19.24	240.33	7.49	7.98	9.07	7.53	29.04	54.52	154.34	17.75	16.72	10.53	14.58	r101_c4_3x
	13.20	19.37	237.53	7.52	7.98	8.98	7.53	28.50	53.64	150.67	17.37	15.91	10.18	13.74	r50_fpn_3x
	13.02	20.89	238.27	7.35	7.91	8.98	7.45	28.32	52.55	150.76	17.21	15.87	10.12	13.71	r101_fpn_3x
	12.98	19.27	236.44	7.32	7.87	8.93	7.41	27.12	51.27	146.81	16.12	15.92	10.00	13.76	x101_fpn_3x
	12.65	19.92	233.05	7.01	7.64	8.80	7.18	24.72	44.25	137.65	13.83	14.83	9.60	12.68	r50_fpn_400e
	12.55	19.59	231.94	6.93	7.58	8.73	7.12	24.99	41.32	139.09	14.02	14.32	9.38	12.15	regnety_400e
	12.60	19.04	231.50	6.96	7.65	8.78	7.19	23.64	39.60	134.20	12.69	14.12	9.27	11.94	regnetx_400e
	12.69	19.01	232.26	7.05	7.69	8.78	7.23	23.16	40.47	132.55	12.25	13.67	9.17	11.49	r101_fpn_400e
	12.93	19.54	235.95	7.26	7.82	8.91	7.37	27.17	49.52	146.71	16.09	15.58	9.94	13.42	**mean**
	0.28	0.99	3.44	0.25	0.17	0.13	0.17	3.34	7.95	11.60	3.16	1.39	0.59	1.41	**std**

Table 13. **The effectiveness of proposed temporal consistency modules on V-HIM60** (Extension of Table 6). The combination of bi-directional Conv-GRU and forward-backward fusion achieves the best overall performance on three test sets. **Bold** highlights the best for each level.

Conv-GRU		Fusion		MAD	MAD_f	MAD_u	MSE	SAD	Grad	Conn	dtSSD	MESSDdt
Single	Bi	$\hat{\mathbf{A}}^f$	$\hat{\mathbf{A}}^b$									
Easy level												
				10.26	13.64	192.97	4.08	3.73	4.12	3.47	16.57	16.55
✓				10.15	12.83	192.69	4.03	3.71	4.09	3.44	16.42	16.44
	✓			10.14	12.70	192.67	4.05	3.70	4.09	3.44	16.41	16.42
	✓	✓		11.32	20.13	194.27	5.01	4.10	4.67	3.85	16.51	17.85
	✓	✓	✓	**10.12**	**12.60**	**192.63**	**4.02**	**3.68**	**4.08**	**3.43**	**16.40**	**16.41**
Medium level												
				13.88	4.78	202.20	5.27	5.56	6.30	**5.11**	23.67	38.90
✓				13.84	4.56	202.13	5.44	5.70	6.35	5.14	23.66	38.25
	✓			**13.83**	4.52	**202.02**	5.39	5.63	6.33	5.12	23.66	38.22
	✓	✓		15.33	9.02	207.61	6.45	6.09	7.56	5.64	24.08	39.82
	✓	✓	✓	13.85	**4.48**	**202.02**	**5.37**	**5.53**	**6.31**	**5.11**	**23.63**	**38.12**
Hard level												
				21.62	30.06	253.94	11.69	7.38	7.07	7.01	30.50	43.54
✓				21.26	28.60	253.42	**11.46**	7.25	7.12	6.95	29.95	43.03
	✓			21.25	28.55	253.17	11.56	7.25	7.10	6.91	29.92	43.01
	✓	✓		24.97	45.62	260.08	14.62	8.55	9.92	8.17	30.66	48.03
	✓	✓	✓	**21.23**	**28.49**	**252.87**	11.53	**7.24**	**7.08**	**6.89**	**29.90**	**42.98**

9. Video matting

This section elaborates on the video matting aspect of our work, providing details about dataset generation and offering additional quantitative and qualitative analyses. For an enhanced viewing experience, we recommend visit our website, which contains video samples from V-HIM60 and real video results of our method compared to baseline approaches.

9.1. Dataset generation

To create our video matte dataset, we utilized the BG20K dataset for backgrounds and incorporated video backgrounds from VM108. We allocated 88 videos for training and 20 for testing, ensuring each video was limited to 30 frames. To maintain realism, each instance within a video displayed an equal number of randomly selected frames from the source videos, with their sizes adjusted to fit within the background height without excessive overlap.

We categorized the dataset into three levels of difficulty, based on the extent of instance overlap:

- **Easy Level:** Features 2-3 distinct instances per video with no overlap.
- **Medium Level:** Includes up to 5 instances per video, with occlusion per frame ranging from 5 to 50%.
- **Hard Level:** Also comprises up to 5 instances but with a higher occlusion range of 50 to 85%, presenting more complex instance interactions.

Table 14. **Our framework outperforms baselines in almost metrics on V-HIM60** (Extension of Table 7). We extend the result in the main paper with more metrics and our model is the best overall. **Bold** and underline indicates the best and second-best model among baselines in the same test set.

Model	MAD	MAD_f	MAD_u	MSE	SAD	Grad	Conn	dtSSD	MESSDdt
Easy level									
MGM-TCVOM	11.36	18.49	202.28	5.13	4.11	4.57	3.83	17.02	19.69
MGM*-TCVOM	10.97	20.33	**187.59**	5.04	3.98	4.19	3.70	16.86	**15.63**
InstMatt	13.77	38.17	219.00	5.32	4.96	4.95	3.98	17.86	18.22
SparseMat	12.02	21.00	205.41	6.31	4.37	4.49	4.11	19.86	24.75
Ours	**10.12**	**12.60**	192.63	**4.02**	**3.68**	**4.08**	**3.43**	**16.40**	16.41
Medium level									
MGM-TCVOM	14.76	4.92	218.18	5.85	5.86	7.17	5.41	**23.39**	39.22
MGM*-TCVOM	**13.76**	4.61	**201.58**	5.50	**5.49**	6.47	**5.02**	23.99	42.71
InstMatt	19.34	35.05	223.39	7.50	7.55	7.21	6.02	24.98	54.27
SparseMat	18.20	10.59	250.89	10.06	7.30	8.03	6.87	30.19	85.79
Ours	13.85	**4.48**	202.02	**5.37**	5.53	**6.31**	5.11	23.63	**38.12**
Hard level									
MGM-TCVOM	22.16	31.89	271.27	11.80	7.65	7.91	7.27	31.00	47.82
MGM*-TCVOM	22.59	36.01	264.31	13.03	7.75	7.86	7.32	32.75	**37.83**
InstMatt	27.24	58.23	275.07	14.40	9.23	7.88	8.02	31.89	47.19
SparseMat	24.83	32.26	312.22	15.87	8.53	8.47	8.19	36.92	55.98
Ours	**21.23**	**28.49**	**252.87**	**11.53**	**7.24**	**7.08**	**6.89**	**29.90**	42.98

During training, we applied dilation and erosion kernels to binarized alpha mattes to generate input masks. For testing purposes, masks were created using the XMem technique, based on the first-frame binarized alpha matte.

We have prepared examples from the testing dataset across all three difficulty levels, which can be viewed in the website for a more immersive experience. The datasets V-HIM2K5 and V-HIM60 will be made publicly available following the acceptance of this work.

9.2. Training details

For video dataset training (V-HIM2K5), we initialized our model with weights from the image pretraining phase. The training involved processing approximately 2.5M frames, using a batch size of 4 and a frame sequence length of $T = 5$ on 8 A100 GPUs. We adjusted the learning rate to 5×10^{-5}, maintaining the cosine learning rate decay with a 1,000-iteration warm-up. In addition to the image augmentations, we incorporated motion blur (from OTVM) during training. Image sizes are kept the same as previously. The first 3,000 iterations continued to use curriculum learning. In addition to the image augmentations, we incorporated motion blur (from OTVM) during training. For testing, the frame size was standardized to a short-side length of 576 pixels.

9.3. Quantitative details

Our ablation study, detailed in Table 13, focuses on various temporal consistency components. The results demonstrate that our proposed combination of Bi-Conv-GRU and forward-backward fusion outperforms other configurations across all metrics. Additionally, Table 14 compares our model's performance against previous baselines using various error metrics. Our model consistently achieves the lowest error rates in almost all metrics.

An illustrative comparison of the impact of different temporal modules is presented in Fig. 18. The addition of Conv-GRU significantly reduces noise, although some residual noise remains. Implementing forward fusion $\hat{\mathbf{A}}^f$ enhances temporal consistency but also propagates errors from previous frames. This issue is effectively addressed by integrating $\hat{\mathbf{A}}^b$, which balances and corrects these errors, improving overall performance.

In an additional experiment, we evaluated trimap-propagation matting models (OTVM [45], FTP-VM [17]), which typically receive a trimap for the first frame and propagate it through the remaining frames. To make a fair comparison with our approach, which utilizes instance masks for each frame, we integrated our model with these trimap-propagation models. The trimap predictions were binarized and used as input for our model. The results, as shown in Table 15, indicate a significant improvement in accuracy when our model is used, compared to the original matte decoder of the trimap-propagation models. This experiment underscores the flexibility and robustness of our proposed framework, which is capable of handling various mask qualities and mask generation methods.

9.4. More qualitative results

For a more immersive and detailed understanding of our model's performance, we recommend viewing the examples on our website which includes comprehensive results and comparisons with previous methods. Additionally, we have highlighted outputs from specific frames in Fig. 19.

Regarding temporal consistency, SparseMat and our framework exhibit comparable results, but our model demonstrates more accurate outcomes. Notably, our output maintains a level of detail on par with InstMatt, while ensuring consistent alpha values across the video, particularly in background and foreground regions. This balance between detail preservation and temporal consistency highlights the advanced capabilities of our model in handling the complexities of video instance matting.

For each example, the first-frame human masks are generated by r101_fpn_400e and propagated by XMem for the rest of the video.

Table 15. **Our framework also reduces the errors of trimap propagation baselines.** When replacing those models' matte decoders with ours, the number in all error metrics was reduced by a large margin. Gray rows denote the module from public weights without retraining on our V-HIM2K5 dataset.

Trimap prediction	Matte decoder	MAD	MAD$_f$	MAD$_u$	MSE	SAD	Grad	Conn	dtSSD	MESSDdt
Easy level										
OTVM	OTVM	204.59	6.65	208.06	192.00	76.90	15.25	76.36	46.58	397.59
OTVM	OTVM	36.56	299.66	382.45	29.08	14.16	6.62	14.01	24.86	69.26
OTVM	Ours	31.00	260.25	326.53	24.58	12.15	5.76	11.94	22.43	55.19
FTP-VM	FTP-VM	12.69	9.13	233.71	5.37	4.66	6.03	4.27	19.83	18.77
FTP-VM	FTP-VM	13.69	24.54	269.88	6.12	5.07	6.69	4.78	20.51	22.54
FTP-VM	Ours	9.03	4.77	194.14	3.07	3.31	3.94	3.08	16.41	15.01
Medium level										
OTVM	OTVM	247.97	14.20	345.86	230.91	98.51	21.02	97.74	66.09	587.47
OTVM	OTVM	48.59	275.62	416.63	37.29	17.25	10.19	17.03	36.06	80.38
OTVM	Ours	36.84	209.77	333.61	27.52	13.04	8.63	12.69	32.95	70.84
FTP-VM	FTP-VM	40.46	32.59	287.53	28.14	15.80	12.18	15.13	32.96	125.73
FTP-VM	FTP-VM	26.86	28.73	318.43	15.57	10.52	12.39	9.95	32.64	126.14
FTP-VM	Ours	18.34	11.02	234.39	9.39	6.97	6.83	6.59	26.39	50.31
Hard level										
OTVM	OTVM	412.41	231.38	777.06	389.68	146.76	29.97	146.11	90.15	764.36
OTVM	OTVM	140.96	1243.20	903.79	126.29	47.98	17.60	47.84	59.66	298.46
OTVM	Ours	123.01	1083.71	746.38	111.16	41.52	16.41	41.24	55.78	257.28
FTP-VM	FTP-VM	46.77	66.52	399.55	33.72	16.33	14.40	15.82	45.04	76.48
FTP-VM	FTP-VM	48.11	95.17	459.16	35.56	16.51	14.87	16.12	45.29	78.66
FTP-VM	Ours	30.12	62.55	326.61	19.13	10.37	8.61	10.07	36.81	66.49

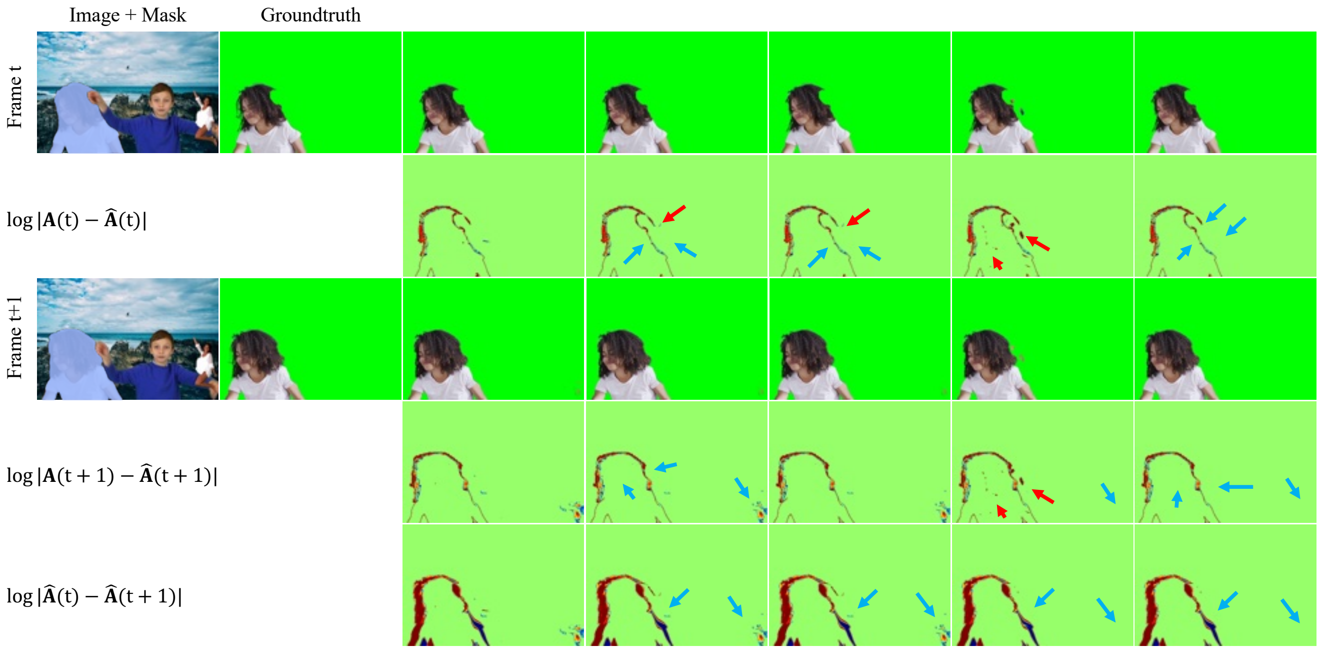
Image + Mask
Groundtruth
Frame t
$\log|\mathbf{A}(t) - \hat{\mathbf{A}}(t)|$
Frame t+1
$\log|\mathbf{A}(t+1) - \hat{\mathbf{A}}(t+1)|$
$\log|\hat{\mathbf{A}}(t) - \hat{\mathbf{A}}(t+1)|$

Conv-GRU	Single		✓			
	Bidirectional			✓	✓	✓
Fusion	$\hat{\mathbf{A}}^f$				✓	✓
	$\hat{\mathbf{A}}^b$					✓

Figure 18. **The effectiveness of different temporal components on the medium level of V-HIM60.** Conv-GRU can improve the result a bit, but not perfect. Our proposed fusion strategy improves the output in both foreground and background regions. The table below denotes temporal components for each column. Red, blue arrows indicate the errors and improvements in comparison with the result without any temporal module. We also visualize the error to the groundtruth (log $|\mathbf{A} - \hat{\mathbf{A}}|$) and the difference between consecutive predictions(log $|\hat{\mathbf{A}} - \hat{\mathbf{A}}|$). The color-coded map (min-max range) to illustrate differences between consecutive frames is . (Best viewed in color and digital zoom).

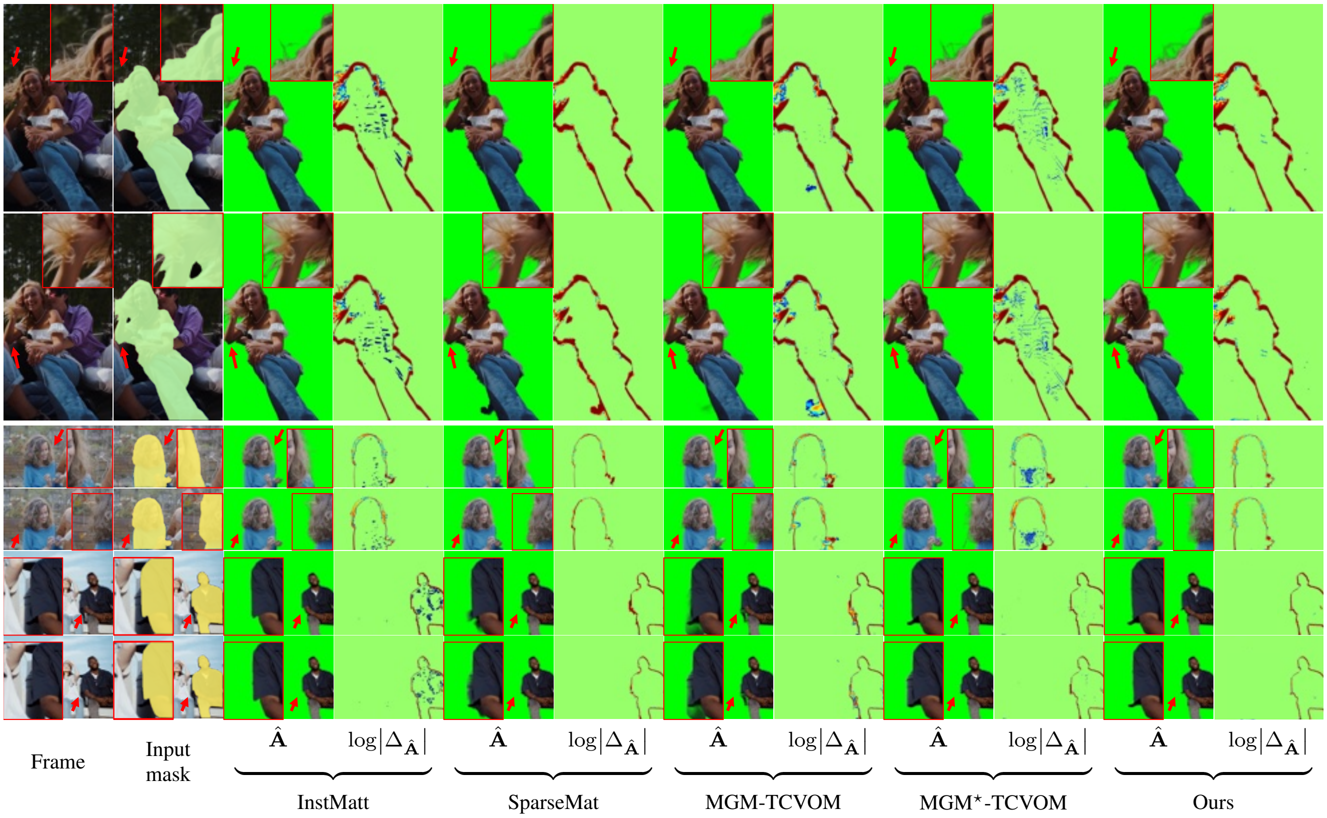

Figure 19. **Highlighted detail and consistency on natural video outputs**. To watch the full videos, please check our website. We present the foreground extracted and the difference to the previous frame output for each model. The color-coded map (min-max range) to illustrate differences between consecutive frames is . Red arrows indicate the zoom-in region in the red square. (Best viewed in color and digital zoom).