SAM2Matting: Generalized Image and Video Matting

Ruiqi Shen[1*], Guangquan Jie[1*], Chang Liu[2✉], and Henghui Ding[1] ✉

[1] Institute of Big Data, College of Computer Science and Artificial Intelligence, Fudan University, China
[2] Shanghai University of Finance and Economics, China
hhding@fudan.edu.cn
https://henghuiding.com/SAM2Matting/

Abstract. Despite impressive advances in image matting, video matting remains challenging due to the inherent gap between high-level tracking, which requires frame-wise understanding, and low-level matting, which focuses on extremely fine-grained details. Existing methods attempt this using costly domain-specific video matting datasets, which may limit their out-of-domain generalization and leave tracking robustness vulnerable. We rethink this paradigm with *SAM2Matting*, a novel framework that decouples the task by enhancing the foundational tracker of SAM2 with a region-proposal bridge and dedicated matting heads. This enables uncompromised SAM2 to handle tracking while the matting components focus exclusively on resolving fine-grained intricate details. Notably, despite being trained only on images, SAM2Matting establishes new state-of-the-art performance on video matting, with robust generalization across both human-centric and in-the-wild matting scenarios.

Keywords: Video Matting · Image Matting · Real-time Matting

1 Introduction

Matting aims to separate the foreground target from its background by predicting a pixel-level alpha matte. Being evaluated by transparency values, it is generally considered as a fundamental low-level vision task [27, 29, 50, 54].

When extending to videos, it is common practice to require user input, typically an initial-frame mask, to disambiguate the target and enable consistent tracking across subsequent frames [17, 37, 52, 59]. Consequently, video matting faces a fundamental trade-off: it demands both high-level semantic understanding to robustly track the target as in Video Object Segmentation (VOS) [38] and low-level fine-grained perception to capture extremely intricate details as in image matting. To bridge this gap, recent approaches rely heavily on video matting datasets [16, 17, 27, 48, 52, 59]. However, the prohibitive cost of annotating pixel-level alpha values across frames restricts these datasets to limited

⋆ Equal contribution
✉ Corresponding authors

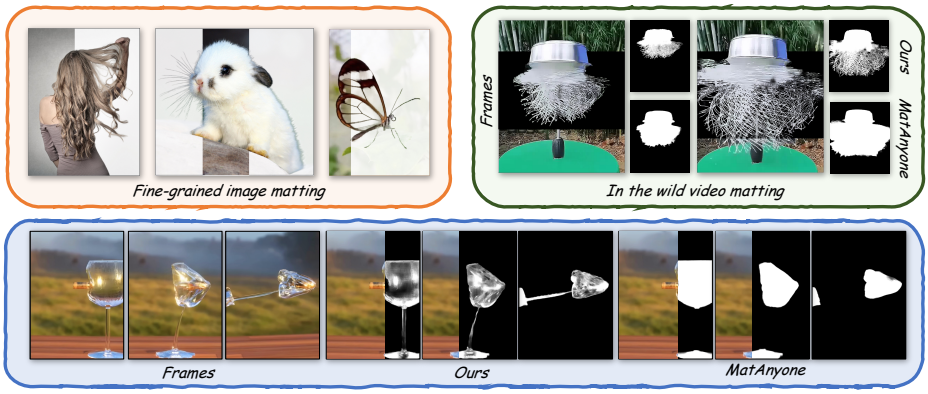

Fig. 1: SAM2Matting achieves robust, high-fidelity image and video matting across both human-centric and in-the-wild scenarios. (Zoom in for details)

scales and narrowly-focused domains, primarily human-centric [17, 27, 52], leaving them insufficient for representing rich real-world dynamics compared with large-scale video segmentation datasets [7–10, 40]. As a result, training a model from scratch on such constrained data fails to establish robust tracking capabilities, while fine-tuning a pretrained VOS model on them compromises its original tracking robustness. These challenges motivate our core question: *must we rely on these heavily-annotated and still narrowly-scoped video matting datasets?*

Rethinking the paradigm, we argue that video matting inherently combines two distinct sub-tasks: high-level tracking for temporal consistency, which is already well-addressed by established VOS models trained at scale, and low-level matting for intricate spatial details, which is comprehensively captured by diverse image matting datasets. We therefore propose *SAM2Matting*, a decoupled framework that preserves the high-level robustness of established VOS tracker while training dedicated components on rich and diverse image matting data for low-level alpha estimation. Our framework seamlessly integrates robust tracking, matte detection, and alpha refinement into a unified pipeline: multi-source priors provide guidance to identify matting-critical regions, where alpha mattes are generated and progressively refined for fine-grained estimation.

Specifically, for each frame, our framework takes multi-scale features, the image, and the SAM2 mask as inputs. To bridge high-level tracking to fine-grained matting, we propose an ROI Detector that identifies matting-critical regions, i.e. regions with fine-grained details or semi-transparency. This differs from prior methods, which rely on rule-based morphological operations for identifying these regions [42, 54, 62] or directly use the raw mask for matting [20, 37, 52, 56]. Subsequently, a Progressive Alpha Predictor generates and refines the mattes within the identified ROIs through a multi-scale cascade, supervised at all intermediate scales [2]. Tailored losses are introduced to smooth transitions and preserve matte integrity, ensuring high-fidelity results.

Experiments show that SAM2Matting achieves new state-of-the-art (SOTA) performance on both image and video matting, with video matting achieved in

a strictly **zero-shot** manner. Extensive in-the-wild results, with representative examples shown in Figure 1, demonstrate its strong generalization to open-world scenarios with rapid motion, complex backgrounds, and target attachments (e.g., man riding a bicycle). Moreover, it is computationally efficient, achieving real-time speed (> 27 FPS) with modest GPU memory. Our contributions are summarized as follows:

- We present *SAM2Matting*, a generalized matting paradigm that decouples video matting into high-level tracking and low-level matting, with each specializing in its task, resulting in optimal performance when combined.
- We integrate diverse priors to identify matting-critical regions, followed by progressive alpha flows for cascaded alpha refinement. Our framework can adapt to different VOS trackers and is robust to tracking inaccuracies.
- SAM2Matting achieves new SOTA performance on video matting in a *zero-shot* way, eliminating costly annotations while generalizing robustly to both human and challenging in-the-wild scenarios.

2 Related Work

2.1 Image Matting

Given an image I, matting aims to separate a foreground target F from its background B with pixel-level precision by predicting an alpha matte α , letting $I = \alpha F + (1 - \alpha)B$. Previous deep image matting methods can be broadly categorized into two paradigms. Automatic (or auxiliary-free) matting directly mats all objects within the image [5,6,22,57,60], requiring no additional input beyond the image itself. However, this paradigm struggles in real-world complex scenes where target identification becomes ambiguous [17, 52]. In contrast, prompt-based matting requires the target to be specified, with earlier methods relying on trimaps [14, 15, 26, 36, 55] for precise guidance. Recently, some methods have relaxed the trimap requirement to coarser forms such as points, scribbles, boxes, or directly use masks that are derived from image or multi-modal segmentation methods, achieving promising results [11, 20, 30, 31, 33, 37, 46, 49, 53, 54, 58].

2.2 Video Matting

A few early video matting approaches [4,19,21,28] are target-agnostic, estimating alpha mattes for all visible objects across the sequence. However, this setting becomes ambiguous and under-defined in real-world where targets frequently enter and exit the frame [16, 17, 52]. Recent methods therefore require explicit target specification: earlier approaches use per-frame or initial-frame trimaps [41,45,61], while newer ones replace trimaps with masks [16, 48, 52, 59] and adapt VOS priors through training on video matting data. However, their performance remains limited by the scale and diversity of these video matting datasets, which require exhaustive annotations and are largely human-centric [17, 27]. Recent methods partially automate alpha-matte generation through green-screen keying, but still rely on human labour to filter artifacts and ensure annotation quality [52]. This motivates us to ask whether robust and generalizable video matting across both human-centric and in-the-wild scenarios can be achieved zero-shot.

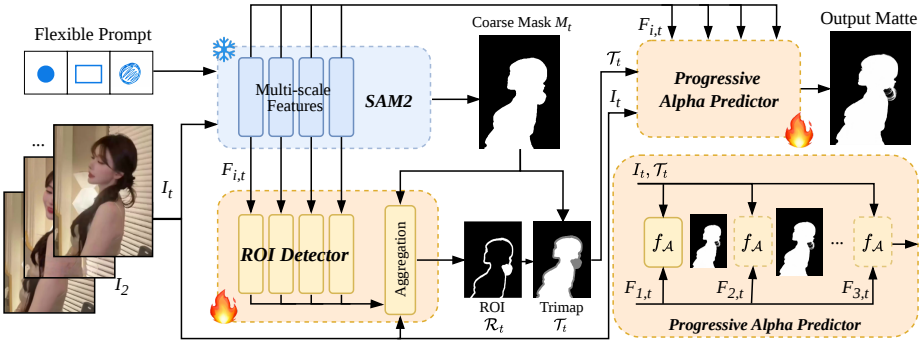

Fig. 2: Overview of SAM2Matting. SAM2Matting adopts a decoupled design that leverages the SAM2 backbone for high-level tracking and dedicated components for low-level matting. The Region of Interest (ROI) Detector identifies matting-critical regions by integrating image-level and mask-level priors. The Progressive Alpha Predictor then generates and refines the alpha mattes across scales through cascaded refinement.

3 Methodology

3.1 Overview

Figure 2 illustrates SAM2Matting, a generalized framework for image and video matting that decouples robust high-level tracking from dedicated low-level matting. Specifically at each frame, SAM2 provides multiple priors, including masks and multi-scale image features, from which a context-aware ROI Detector identifies matting-critical regions with fine-grained details or semi-transparency. A Progressive Alpha Predictor then iteratively produces and refines the matte through a coarse-to-fine cascade, with intermediate mattes supervised at each scale to progressively capture finer details.

3.2 Regions of Interest (ROI) Detector

Typical matting frameworks generate regions of interest (ROI), or "unknown" regions, as an intermediate step to concentrate the model on the most relevant areas for alpha estimation. However, conventional approaches derive these regions in simple rule-based ways. One common strategy employs morphological operations on the mask (e.g. dilation and erosion) [42, 54, 62], which implicitly assumes uniform boundary importance (Figure 3). Alternatively, other methods directly use the raw mask as the ROI [20, 37, 56]. Both approaches are prone to overlook fine details within complex structures or include definite foreground regions that do not require matting.

To overcome these limitations, we introduce an ROI Detector to precisely identify matting-relevant regions in each frame. We reformulate ROI detection as a pixel-wise binary classification task, where positive pixels denote matting-critical areas. The detector integrates diverse priors, including the SAM2 mask

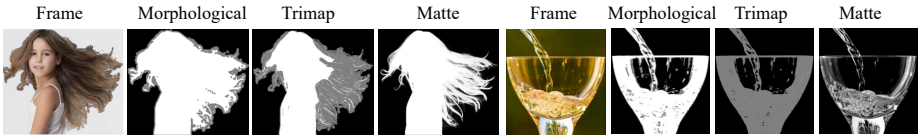

Fig. 3: Simple morphological operations (dilation & erosion) produce coarse, uniform boundaries ("Morphological"), which fail to capture complex regions in the ground-truth trimap ("Trimap"), such as flying hair (*left*) and translucent water cup (*right*).

M, the current frame I, and the multi-scale image features F. Since features at different resolutions capture different levels of semantics and details, we predict an ROI logit map at each scale $i \in \{1, \ldots, n\}$.

Specifically, at frame t, for each scale $i \in \{1, \ldots, n\}$, a convolutional head $f_{\mathrm{R},i}(\cdot)$ estimates an ROI logit map $L_{t,i} \in \mathbb{R}^{H_i \times W_i}$ from the image feature $F_{t,i} \in \mathbb{R}^{C_i \times H_i \times W_i}$, the resized frame $I_{t,i} \in \mathbb{R}^{3 \times H_i \times W_i}$, and the resized mask $M_{t,i} \in \{0,1\}^{H_i \times W_i}$:

$$L_{t,i} = f_{\mathrm{R}}\left(I_{t,i}, M_{t,i}, F_{t,i}\right). \tag{1}$$

The logit maps at different scales are then aggregated by a hierarchical convolutional network $f_{\varphi}(\cdot)$, which integrates global context with structural details, producing $L_t \in \mathbb{R}^{H \times W}$ for frame t:

$$L_t = f_{\varphi}\left(\left[U(L_{t,1}), U(L_{t,2}), \cdots, U(L_{t,n})\right]\right), \tag{2}$$

where $U(\cdot)$ and $[\cdot]$ denote upsampling and concatenation, respectively. The final ROI prediction $\mathcal{R}_t \in \{0,1\}^{H \times W}$ for frame t is then obtained by applying a sigmoid activation to L_t and thresholding it with θ:

$$\mathcal{R}_t = \mathbb{1}\left[\sigma(L_t) \geq \theta\right]. \tag{3}$$

During training, we supervise L_t with focal and smoothness losses (see Section 3.5), encouraging the predicted $\mathcal{R}_t$ to precisely capture regions with intricate details or semi-transparency that require matting.

3.3 Pseudo-Trimap Generation

The ROI $\mathcal{R}_t$ identifies regions requiring fine-grained matting. To preserve structural integrity and provide explicit foreground-background separation, we construct a pseudo-trimap $\mathcal{T}_t \in \{0, 0.5, 1\}^{H \times W}$ by assigning the SAM2 mask $M_t \in \{0,1\}^{H \times W}$ to definite foreground and background, while designating $\mathcal{R}_t$ as the unknown region. Specifically, the pseudo-trimap $\mathcal{T}_t$ for frame t is defined as:

$$\mathcal{T}_{t,(h,w)} = \begin{cases} M_{t,(h,w)}, & \text{if } \mathcal{R}_{t,(h,w)} = 0, \\ 0.5, & \text{if } \mathcal{R}_{t,(h,w)} = 1. \end{cases} \tag{4}$$

The resulting $\mathcal{T}_t$ assigns labels for definite foreground (1), definite background (0), and unknown regions (0.5), providing pixel-level spatial priors for subsequent alpha estimation.

3.4 Progressive Alpha Predictor

Unlike the ROI Detector's parallel processing, the Progressive Alpha Predictor treats alpha estimation as a sequential refinement process. It adopts a coarse-to-fine strategy, where each intermediate alpha prediction is passed to the subsequent scale as guidance for refinement. Specifically, at frame t, the composite input $X_{t,i}$ at scale i concatenates the image feature $F_{t,i}$, the resized frame $I_{t,i}$, the resized pseudo-trimap $\mathcal{T}_{t,i}$, and the upsampled matte $\mathcal{A}_{t,i-1}$ from the preceding scale (for $i \geq 2$):

$$X_{t,i} = \begin{cases} [F_{t,1}, \mathcal{T}_{t,1}, I_{t,1}], & i = 1, \\ [F_{t,i}, \mathcal{T}_{t,i}, I_{t,i}, U(\mathcal{A}_{t,i-1})], & i = 2, \ldots, n. \end{cases} \tag{5}$$

A scale-specific projection layer $g_{\mathcal{A},i}(\cdot)$ first maps $X_{t,i}$ to a fixed-dimensional embedding, from which a convolutional matting head $f_{\mathcal{A},i}(\cdot)$ predicts $\mathcal{A}_{t,i}$ at scale i:

$$\mathcal{A}_{t,i} = \sigma(f_{\mathcal{A},i}(g_{\mathcal{A},i}(X_{t,i}))), \quad \mathcal{A}_{t,i} \in (0,1)^{H_i \times W_i}. \tag{6}$$

Finally, the matte $\mathcal{A}_{t,n}$ at the finest scale $i = n$ is upsampled to the original resolution, yielding the final alpha matte $\mathcal{A}_t \in (0,1)^{H \times W}$ for frame t.

3.5 Optimization Strategies

Training Designs. We freeze the SAM2 backbone to preserve its high-level tracking capability for superior temporal consistency, while training only the matting components on high-quality image matting data, enabling fine-grained alpha refinement without compromising tracker robustness.

Loss Designs. Since SAM2Matting is trained solely on static images, supervision operates per frame. For a training sample at index t, we first threshold the ground-truth alpha matte $\mathcal{A}_{\mathrm{GT}_t}$ within the range $[\alpha, \beta]$, followed by dilation and erosion to obtain the ground-truth ROI $\mathcal{R}_{\mathrm{GT}_t}$. The ROI Detector is then optimized with a focal loss $\mathcal{L}_{\mathrm{focal}}$ for pixel-wise binary classification and a smooth-L_1 loss $\mathcal{L}_{\mathrm{smooth}}$ to reduce jagged artifacts common at ROI boundaries [18, 48, 52]:

$$\delta_t = \mathbf{1}\{\alpha \leq \mathcal{A}_{\mathrm{GT}_t} \leq \beta\}, \quad \mathcal{R}_{\mathrm{GT}_t} = \mathrm{Dilate}(\delta_t) - \mathrm{Erode}(\delta_t). \tag{7}$$

$$\mathcal{L}_{\mathcal{R}} = \mathcal{L}_{\mathrm{focal}}(L_t, \mathcal{R}_{\mathrm{GT}_t}) + \lambda_s \mathcal{L}_{\mathrm{smooth}}(L_t, \mathcal{R}_{\mathrm{GT}_t}). \tag{8}$$

Following [14, 20, 28], we adopt L_1 loss $\mathcal{L}_{L_1}$ and Laplacian loss $\mathcal{L}_{\mathrm{lap}}$ for fine-grained alpha estimation. Inspired by the auxiliary loss in [2], we apply deep supervision across all prediction scales $i \in \{1, \ldots, n\}$. Furthermore, to preserve matte integrity and prevent hollow regions inside the matte foreground [14, 17, 19, 21, 34], we apply a matte-mask consistency penalty $\mathcal{L}_{\mathrm{con}}$. Specifically, we threshold the predicted matte $\mathcal{A}_t$ at γ to produce a binary mask $M_{\mathcal{A}_t}$, which is then anchored to the SAM2 mask M_t as follows:

$$M_{\mathcal{A}_t}(h,w) = \begin{cases} 1, & \mathcal{A}_t(h,w) > \gamma, \\ 0, & \text{otherwise}, \end{cases} \tag{9}$$

$$\mathcal{L}_{\text{con}} = \mathcal{L}_{\text{seg}}\big(M_{\mathcal{A}_t},\, M_t\big), \tag{10}$$

where $\mathcal{L}_{\text{seg}}$ is a joint segmentation loss comprising focal and dice terms. The joint supervision $\mathcal{L}_{\mathcal{A}}$ for the Progressive Alpha Predictor is then given by:

$$\mathcal{L}_{\mathcal{A}} = \mathcal{L}_{L_1} + \mathcal{L}_{\text{lap}} + \mathcal{L}_{\text{con}}. \tag{11}$$

Finally, the overall training objective $\mathcal{L}$ combining ROI detection and alpha refinement is formulated as:

$$\mathcal{L} = \mathcal{L}_{\mathcal{R}} + \mathcal{L}_{\mathcal{A}}. \tag{12}$$

4 Experiments

4.1 Experimental Setup

Training Data. We train a full model for optimal performance using 8 image matting datasets: I-HIM50K [17], P3M-10k [22], CelebAHairMask-HQ [13], AIM-500 [24], Distinctions-646 [39], AM-2K [23], UHRIM [51], and RefMatte [25]. To align with existing baselines, we also evaluate variants trained on the same datasets as those baselines (see Section 4.2), allowing us to better examine the role of our architectural design.

Metrics. Following prior works [17, 50, 52, 54], we evaluate performance using standard matting metrics: Mean Absolute Difference (MAD), Mean Squared Error (MSE), Gradient (Grad), Connectivity (Conn) and dtSSD (video matting only). For all metrics, lower values are better.

Implementation Details. We adopt the SAM2.1 Hiera-B+ [40] as our tracking backbone. Following our decoupled design, the backbone remains frozen with only the matting components optimized. The model is trained for 5 epochs on 4 NVIDIA A6000 GPUs using the AdamW optimizer [35] with a batch size of 32 and a learning rate of 5×10^{-3}. To ensure robust generalization, we employ data augmentation including random horizontal flipping, affine transformations, color jittering, and Gaussian noise. We set $\theta = 0.65$ for the ROI Detector, while α, β, and γ are optimized to 0.15, 0.5, and 0.05 via grid search, respectively.

4.2 Quantitative Evaluation on Image Matting

We evaluate SAM2Matting on three key benchmarks: P3M-500-NP [22], AM-2K [23], and PPM-100 [18], comparing it with previous SOTAs that similarly build on SAM, including MAM [20] and Matte Anything [54]. We also train two variants (*, †) using the same datasets as those baselines for fair comparisons. Table 1 shows that SAM2Matting consistently achieves superior performance: it outperforms MAM with an 11.35 lower MAD on P3M-500-NP, and surpasses Matte Anything with a 24.8% lower MSE on AM-2K, demonstrating that the ROI Detector effectively captures fine-grained details that might be missed by raw masks (MAM) or morphological trimaps (Matte Anything). See the supplementary material for more comparisons.

Table 1: Image matting results. "–": no reported result. The best and second-best results are marked in red and orange, respectively.

		P3M-500-NP				AM-2K test				PPM-100	
Methods	Datasets	MAD↓	MSE↓	Grad↓	Conn↓	MAD↓	MSE↓	Grad↓	Conn↓	MAD↓	MSE↓
P3M [22]	-	6.50	3.50	–	–	–	–	–	–	9.60	5.80
GFM [23]	-	–	5.60	14.80	18.00	5.90	2.40	9.00	9.40	–	–
E2E-HIM [32]	-	5.40	3.00	–	–	–	–	–	–	7.20	4.00
MAM [20]	*	15.40	9.20	14.22	–	10.10	3.50	10.65	–	9.90	4.60
SAM2Matting	*	4.05	1.14	9.58	6.54	5.38	1.91	7.81	8.65	4.61	1.33
Matte Anything [54]	†	–	2.80	17.30	10.00	–	3.30	11.70	10.90	–	–
SAM2Matting	†	3.94	1.11	8.85	6.45	5.95	2.48	8.25	9.60	4.68	1.42
SAM2Matting	Full	3.81	1.00	8.78	5.97	4.90	1.59	7.23	7.75	4.31	1.22

4.3 Quantitative Evaluation on Video Matting

We benchmark SAM2Matting's video matting performance on V-HIM60 [17] and VideoMatte [27] in a *zero-shot* manner, comparing against baselines trained on video matting datasets, including MatAnyone [52], MaGGIe [17], RVM [28], FTP-VM [16], as well as raw SAM2 segmentation. As shown in Table 2, our method consistently outperforms fully supervised baselines across most metrics despite its zero-shot setting. This gain is particularly pronounced on the challenging V-HIM60-Hard, where it surpasses MatAnyone with an 11.89 reduction in MAD. It also achieves the lowest dtSSD, indicating robust temporal consistency. In contrast, raw SAM2 performs poorly across all metrics, while existing baselines lag behind in spatial accuracy (MAD, MSE). These results demonstrate the effectiveness of our decoupled design, which allows tracking and matting to specialize for optimal performance when combined.

4.4 Qualitative Evaluation

Human Matting. SAM2Matting outperforms strong baselines including RVM [28] and MatAnyone [52], delivering superior results on intricate hair-level details and complex transparencies, such as human with glasses. As illustrated in Figure 4, SAM2Matting not only extracts fine details under complex lighting environment, but also filters out undesired occlusion parts of the foreground, such as the green bars in front of the human.

In-the-Wild Videos. As shown in Figure 5, existing video matting methods (MatAnyone [52], MaGGIe [17], RVM [28]) struggle with in-the-wild sequences, such as rapidly growing roots, semi-transparent butterflies, and rapidly falling water drip, due to the conflict between high-level tracking and low-level matting. In contrast, our decoupling strategy mitigates this issue by securing an accurate track from which dedicated matting components extract fine-grained details.

Attachments and Background Distractions. SAM2Matting also performs well at matting target with attached objects, e.g., human with bicycles or ski

Table 2: Video matting results. SAM2Matting is evaluated in a *zero-shot* setting.

Methods	Reference	V-HIM60-Easy MAD↓	MSE↓	Grad↓	Conn↓	dtSSD↓	V-HIM60-Medium MAD↓	MSE↓	Grad↓	Conn↓	dtSSD↓
OTVM [41]	[ECCV'22]	36.56	–	6.62	14.01	24.86	48.59	–	10.19	17.03	36.06
FTP-VM [16]	[CVPR'23]	13.69	–	6.69	4.78	20.51	26.86	–	12.39	9.95	32.64
InstMatt [43]	[CVPR'22]	13.77	–	4.95	3.98	17.86	19.34	–	7.21	6.02	24.98
SparseMat [44]	[CVPR'23]	12.02	–	4.49	4.11	19.86	18.20	–	8.03	6.87	30.19
MaGGIe [17]	[CVPR'24]	10.12	–	4.08	3.43	16.40	13.85	–	6.31	5.11	23.63
MatAnyone [52]	[CVPR'25]	14.38	7.44	5.28	5.19	3.80	29.95	19.72	9.03	12.28	5.98
SAM2 [40]	[ICLR'25]	21.33	14.91	27.16	8.02	7.42	28.37	19.24	35.37	11.53	8.78
SAM2Matting	–	8.72	2.45	4.43	2.91	3.13	13.71	4.74	7.28	4.89	4.24

Methods	Reference	V-HIM60-Hard MAD↓	MSE↓	Grad↓	Conn↓	dtSSD↓	VideoMatte-SD MAD↓	MSE↓	Grad↓	Conn↓	dtSSD↓
OTVM [41]	[ECCV'22]	140.96	–	17.60	47.84	59.66	–	–	–	–	–
FTP-VM [16]	[CVPR'23]	48.11	–	14.87	16.12	45.29	6.13	1.31	1.14	0.41	1.60
RVM [28]	[WACV'22]	–	–	–	–	–	6.08	1.47	0.88	0.41	1.36
InstMatt [43]	[CVPR'22]	27.24	–	7.88	8.02	31.89	–	–	–	–	–
SparseMat [44]	[CVPR'23]	24.83	–	8.47	8.19	36.92	–	–	–	–	–
MaGGIe [17]	[CVPR'24]	21.23	–	7.08	6.89	29.90	5.49	0.60	0.57	0.31	1.39
MatAnyone [52]	[CVPR'25]	30.09	18.87	8.93	10.00	6.72	5.15	0.93	0.67	0.26	1.18
SAM2 [40]	[ICLR'25]	32.29	22.40	31.72	11.42	9.21	7.75	2.93	2.40	0.69	3.92
SAM2Matting	–	18.20	8.55	7.39	6.01	5.10	4.83	0.36	0.34	0.19	1.15

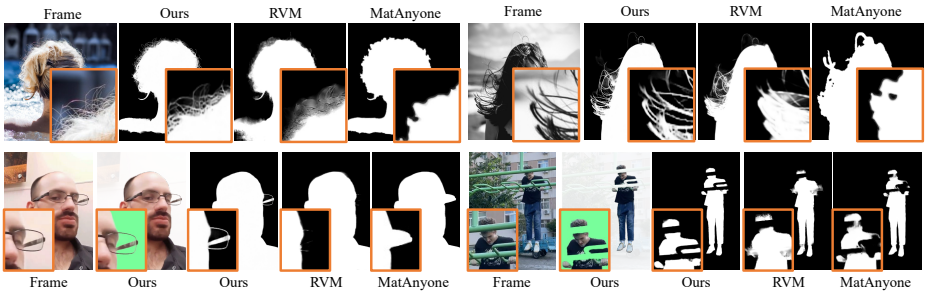

Fig. 4: Qualitative comparisons with RVM [28] and MatAnyone [52] on human matting. SAM2Matting demonstrates superior capabilities in resolving fine-grained details of hair strands and semi-transparencies. (Zoom in for details)

poles, as shown in Figure 6. It also resists subtle background distractions (desk), driven by the matte-mask consistency supervision as outlined in Section 3.5.

4.5 Ablation Studies

ROI Strategies. Table 3 ablates different ROI strategies on V-HIM60-Hard [17]. We compare SAM2Matting against two baselines: (a) "Morphological", which generates ROI using standard dilation and erosion on the mask [42, 54, 62]; and (b) "Full mask", which directly uses the raw SAM2 mask as the ROI [20,37,52,56]. The ROI Detector outperforms both baselines across all metrics, highlighting its ability to robustly identify subtle regions with fine details or semi-transparency.

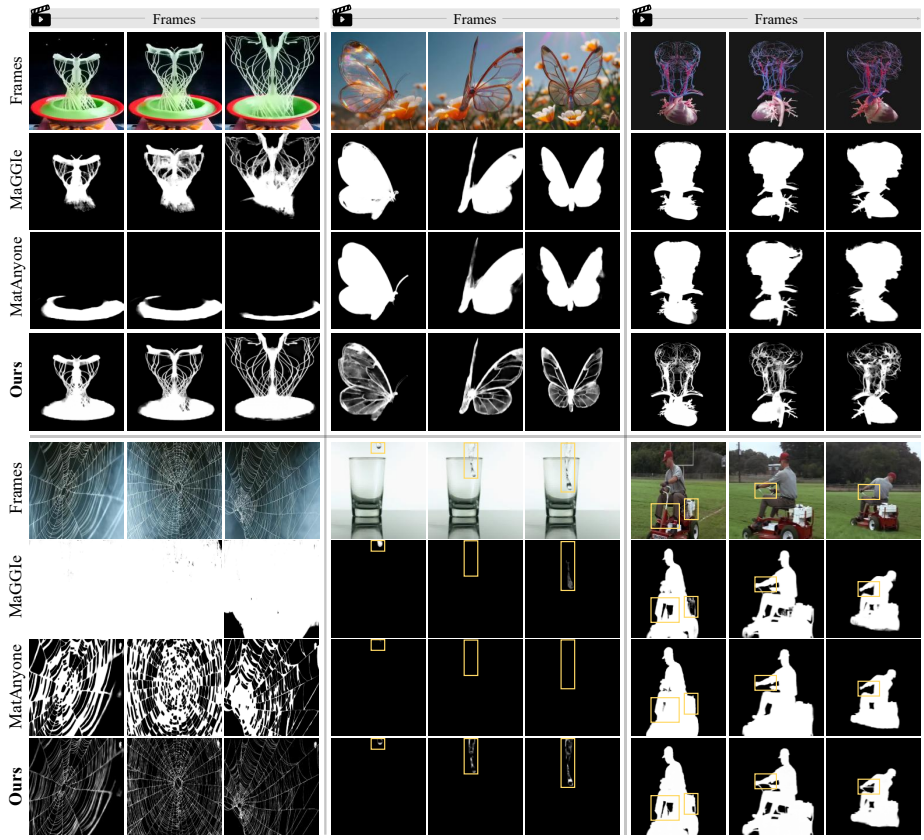

Fig. 5: Qualitative video comparisons with MatAnyone [52] and MaGGIe [17] in challenging in-the-wild scenarios with rapid motion. SAM2Matting exhibits stable tracking and resolves intricate details whereas others fail. (Zoom in for details)

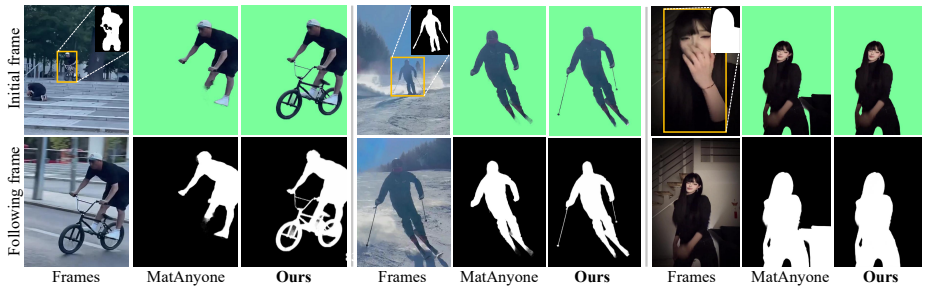

Fig. 6: SAM2Matting robustly preserves target attachments (e.g., bicycle, ski poles) and resists attached background distractors (e.g., desk) in videos. (Zoom in for details)

Figure 7 visualizes this effect, where the ROI Detector captures instance-specific matting-critical regions, such as flying hair, intricate leaves and arm gaps easily overlooked by morphological operations or raw masks.

Table 3: Ablation study on different ROI strategies.

ROI Strategies	MAD ↓	Grad ↓	Conn ↓	dtSSD ↓
(a) Morphological	29.82	11.57	10.37	7.48
(b) Full mask	20.07	9.11	6.68	5.50
(c) **ROI Detector**	**18.20**	**7.39**	**6.01**	**5.10**

Table 4: Ablation study on architectural and supervision designs.

	Prog. Scaling	Con. Loss	Smooth Loss	MAD ↓	Grad ↓	Conn ↓	dtSSD ↓
(a)		✓	✓	19.43	7.88	6.35	5.30
(b)	✓		✓	18.65	7.70	6.20	5.18
(c)	✓	✓		18.26	7.45	6.04	**5.09**
(d)	✓	✓	✓	**18.20**	**7.39**	**6.01**	5.10

Table 5: Ablation study of fine-tuning on large-scale video matting dataset. "Video-FT" denotes the fine-tuned model on V-HIM2K5 [17].

	V-HIM60-Hard			VideoMatte-SD			AM-2K test			PPM-100		
Methods	MAD↓	Grad↓	dtSSD↓	MAD↓	Grad↓	dtSSD↓	MAD↓	Grad↓	Conn↓	MAD↓	Grad↓	Conn↓
Video-FT	**17.90**	**7.08**	**5.01**	4.85	**0.34**	**1.12**	5.23	7.40	7.96	4.40	49.63	38.19
Ours	18.20	7.39	5.10	**4.83**	**0.34**	1.15	**4.90**	**7.23**	**7.75**	**4.31**	**49.17**	**37.58**

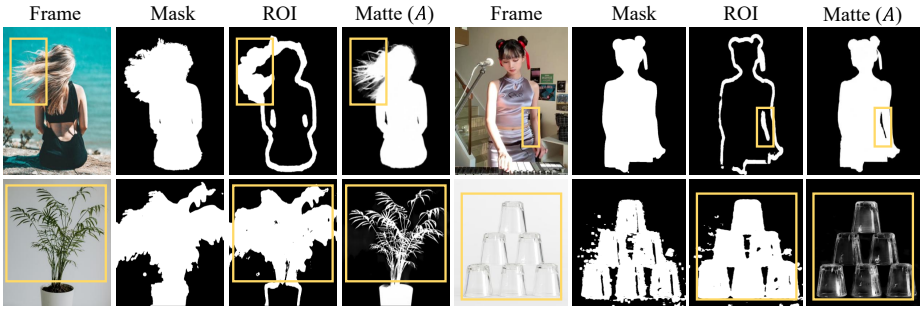

Fig. 7: The ROI Detector identifies target-specific matting-critical regions such as flying hair, leaves of potted plants, and limb gaps that rule-based morphological operations or raw masks fail to capture. (Zoom in for details)

Architecture and Supervision Designs. Table 4 (a) validates the multi-scale cascade refinement in the progressive alpha predictor. Figure 8 shows the predictor exhibiting a coarse-to-fine refinement of alpha mattes (from $\mathcal{A}_1$ to $\mathcal{A}_3$), progressively recovering fine structures such as the hollow chair and flying hair. Table 4 (b,c) further confirms the effectiveness of our supervision designs: the matte-mask consistency penalty (b) fills foreground holes (Figure 9, left), while the smoothness loss (c) reduces jagged boundaries (Figure 9, right). Together, these designs contribute to high-fidelity video matting.

Does fine-tuning on video matting datasets help? We investigate this by fine-tuning our model on V-HIM2K5 [17], a public large-scale human-centric video matting dataset. Table 5 shows a clear trade-off: fine-tuning improves the in-domain benchmark (V-HIM60-Hard) but degrades generalization, with worse results on out-of-domain animal data (AM-2K) and little change on other human-centric datasets (VideoMatte and PPM-100). This indicates overfitting, as the video matting data covers relatively simple scenarios within a narrow domain.

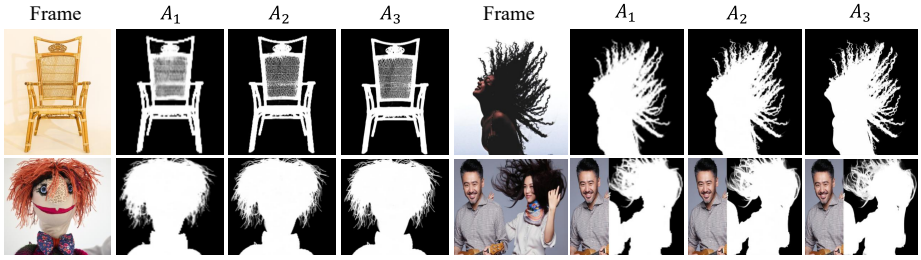

Fig. 8: Intermediate mattes are supervised and iteratively refined across scales, from $\mathcal{A}_1$ to $\mathcal{A}_3$, progressively capturing finer details. (Zoom in for details)

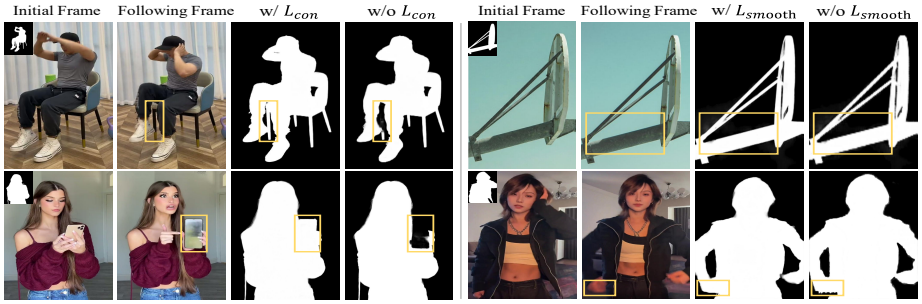

Fig. 9: *Left*: The matte-mask consistency loss L_{con} helps preserve the integrity of alpha mattes. *Right*: The smoothness loss L_{smooth} reduces jagged edges, providing cleaner mattes. (Zoom in for details)

Table 6: Comparison of FPS and VRAM usage (GB) on V-HIM60-Hard [17], measured on a single NVIDIA A6000 GPU. "oom" denotes out-of-memory.

Methods	Res.	FPS↑	VRAM↓	MAD↓	MSE↓	Grad↓	dtSSD↓
MatAnyone 2160p: oom	720p	25.96	3.63	33.49	21.65	9.27	7.04
	1080p	11.29	14.16	30.11	18.89	8.94	6.72
	1440p	2.92	41.31	31.25	19.97	9.96	7.14
SAM2Matting	Any	**27.40**	**3.38**	**18.20**	**8.55**	**7.39**	**5.10**

4.6 FPS and VRAM Efficiency

As shown in Table 6, SAM2Matting supports arbitrary input with stable computational cost and VRAM usage, consistently achieving real-time throughput of >27 FPS. In contrast, MatAnyone [52] experiences a sharp drop in FPS as resolution increases, with only marginal performance gains. Our decoupled approach consistently achieves superior performance across metrics and resolutions.

4.7 Robustness to SAM2 Tracking Inaccuracies.

SAM2Matting mitigates tracking inaccuracies from SAM2. This is because both the ROI Detector and Alpha Predictor leverage fused appearance and shape priors, rather than relying solely on the tracked mask. As shown in Figure 2,

Fig. 10: SAM2Matting robustly mitigates tracking inaccuracies. *Left*: The ROI Detector recovers ski poles missed by SAM2. *Right*: It removes the desk to the woman's right, which is mistakenly included as foreground by SAM2. (Zoom in for details)

Method	Easy	Med.	Hard	VM
FTP-VM	10.4	13.1	18.2	1.6
RVM	12.1	15.2	20.3	1.4
MaGGIe	12.1	15.2	24.2	1.4
MatAnyone	7.5	12.2	17.0	1.3
SAM2Matting	6.8	11.3	15.6	1.0

Table 7: Quantitative results of the flickering effect ($F_{\text{warp}} \times 10^3$)

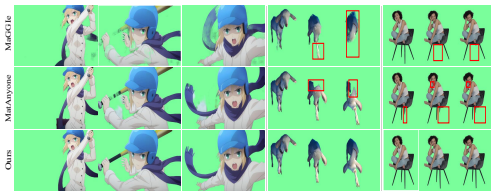

Fig. 11: Qualitative comparison of the flickering effect (Zoom in for details)

the SAM2 mask serves as feature-level guidance rather than a hard constraint, allowing inherent correction of tracking inaccuracies. As illustrated in Figure 10, the multi-source fusion of image features and the SAM2 mask helps the ROI Detector recover the ski poles missed by SAM2 (left). Conversely, when SAM2 incorrectly includes the desk as foreground (right), the ROI Detector effectively identifies it as background and suppresses it, leading to accurate matting results.

4.8 Flickering effect

Flickering reflects temporal inconsistency in video matting. Inspired by [47], we introduce flow-based temporal warping error to quantitatively evaluate this effect:

$$F_{\text{warp}} = \frac{1}{T-1} \sum_{t=2}^{T} \text{MSE}\left(\alpha_t, \mathcal{W}\left(\alpha_{t-1}, F_{t \to t-1}\right)\right), \tag{13}$$

where $\mathcal{W}$ is backward warping and $F_{t \to t-1}$ denotes the optical flow computed by the RAFT Large model [47]. We report F_{warp} on V-HIM60-Easy (Easy), V-HIM60-Medium (Med.), V-HIM60-Hard (Hard), and VideoMatte-SD (VM).

Table 7 shows SAM2Matting consistently outperforms all baselines with the lowest F_{warp}. Figure 11 further visualizes its consistent handling of rapid-changing fine details, whereas previous methods suffer from noticeable flickering.

5 Discussion

Backbone-agnostic nature. Our decoupled framework allows the high-level tracking backbone (i.e., tracker) to be replaced. To show that SAM2Matting does not rely on a specific tracker, we replace the default SAM2.1-B+ with Cutie-base [3], SAM2.1-Tiny (SAM2.1-T) [40], SAM2.1Long-B+ [12], and SAM3 [1]. All variants are trained using the same datasets and strategies in Section 4.1. As shown in Table 8, these variants achieve strong results across video matting benchmarks, suggesting our framework is backbone-agnostic and generalizable.

Table 8: Backbone study across various trackers and benchmarks. SAM2.1-B+ is the default tracker for SAM2Matting.

Tracker	V-HIM60-Easy		V-HIM60-Med.		V-HIM60-Hard		VideoMatte-SD	
	MAD↓	dtSSD↓	MAD↓	dtSSD↓	MAD↓	dtSSD↓	MAD↓	dtSSD↓
Cutie-base [3]	9.45	3.31	15.18	4.62	19.87	5.60	5.79	1.45
SAM2.1-T [40]	9.07	3.20	14.06	4.28	18.43	5.29	5.06	1.23
SAM2.1-B+ [40]	8.72	3.13	13.71	4.24	18.20	5.10	4.83	1.15
SAM2.1Long-B+ [12]	8.70	3.13	13.36	4.12	18.09	5.10	5.10	1.13
SAM3 [1]	7.88	2.89	11.77	3.81	14.37	4.37	4.44	1.11

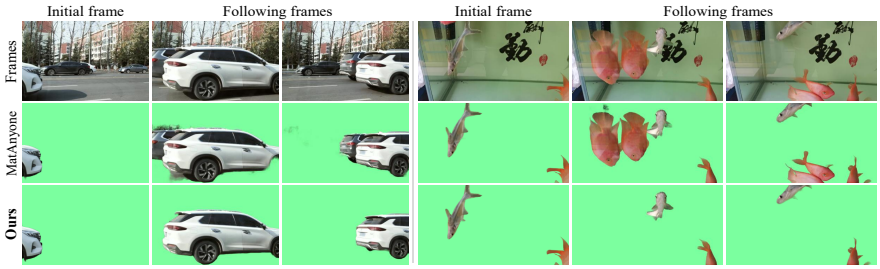

Fig. 12: Comparison of tracking stability on real-world video sequences.

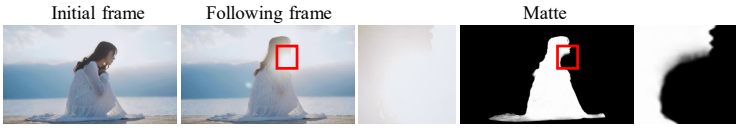

Fig. 13: Failure case under extreme overexposure.

Tracking stability. We compare the tracking stability of SAM2Matting with MatAnyone [52] on real-world videos. As shown in Figure 12, our method produces temporally stable mattes in cases such as moving cars (left) and swimming fish (right). In contrast, MatAnyone suffers from tracking loss and inter-frame inconsistencies in these ordinary cases, suggesting that a strong foundational tracker can provide temporal coherence without explicit temporal modeling.
Failure case. Our approach might degrade in some extreme cases where visual cues are inherently lost due to severe overexposure to lighting and ground-truth becomes undefinable, as shown in Figure 13.

6 Conclusion

We present SAM2Matting, a generalized image and video matting framework that decouples high-level tracking from low-level matting. It achieves state-of-the-art video matting performance without costly video matting datasets, and generalizes to both human-centric and in-the-wild scenarios with temporal consistency. It also accommodates different trackers and interactive prompting. We hope SAM2Matting can support real-world applications and future research.

Acknowledgements. This work was supported by the National Natural Science Foundation of China (NSFC) under Grant No. 62472104 and the Science and Technology Commission of Shanghai Municipality under Grant No. 25511103600.

References

1. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., et al.: Sam 3: Segment anything with concepts. arXiv preprint arXiv:2511.16719 (2025)
2. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1290–1299 (2022)
3. Cheng, H.K., Oh, S.W., Price, B., Lee, J.Y., Schwing, A.: Putting the object back into video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3151–3161 (2024)
4. Chuang, Y.Y., Agarwala, A., Curless, B., Salesin, D.H., Szeliski, R.: Video matting of complex scenes. In: Proceedings of the 29th annual conference on Computer graphics and interactive techniques. pp. 243–248 (2002)
5. Dai, L., Song, X., Liu, X., Li, C., Shi, Z., Chen, J., Brooks, M.: Enabling trimap-free image matting with a frequency-guided saliency-aware network via joint learning. IEEE Transactions on Multimedia **25**, 4868–4879 (2022)
6. Deora, R., Sharma, R., Raj, D.S.S.: Salient image matting. arXiv preprint arXiv:2103.12337 (2021)
7. Ding, H., Liu, C., He, S., Jiang, X., Loy, C.C.: MeViS: A large-scale benchmark for video segmentation with motion expressions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10156–10166 (2023)
8. Ding, H., Liu, C., He, S., Jiang, X., Torr, P.H., Bai, S.: MOSE: A new dataset for video object segmentation in complex scenes. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 20224–20234 (2023)
9. Ding, H., Liu, C., He, S., Ying, K., Jiang, X., Loy, C.C., Jiang, Y.G.: MeViS: A multi-modal dataset for referring motion expression video segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **47**(12), 11400–11416 (2025). https://doi.org/10.1109/TPAMI.2025.3600507
10. Ding, H., Ying, K., Liu, C., He, S., Jiang, X., Jiang, Y.G., Torr, P.H., Bai, S.: MOSEv2: A more challenging dataset for video object segmentation in complex scenes. arXiv preprint arXiv:2508.05630 (2025)
11. Ding, H., Zhang, H., Liu, C., Jiang, X.: Deep interactive image matting with feature propagation. IEEE Transactions on Image Processing **31**, 2421–2432 (2022)
12. Ding, S., Qian, R., Dong, X., Zhang, P., Zang, Y., Cao, Y., Guo, Y., Lin, D., Wang, J.: Sam2long: Enhancing sam 2 for long video segmentation with a training-free memory tree. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13614–13624 (2025)
13. Gao, Z.: Celebahairmask-hq. https://github.com/cpuimage/CelebAHairMask-HQ (2025), accessed: 2026-05-29
14. Hou, Q., Liu, F.: Context-aware image matting for simultaneous foreground and alpha estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4130–4139 (2019)

15. Hu, Y., Lin, Y., Wang, W., Zhao, Y., Wei, Y., Shi, H.: Diffusion for natural image matting. In: European Conference on Computer Vision. pp. 181–199. Springer (2024)
16. Huang, W.L., Lee, M.S.: End-to-end video matting with trimap propagation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14337–14347 (2023)
17. Huynh, C., Oh, S.W., Shrivastava, A., Lee, J.Y.: Maggie: Masked guided gradual human instance matting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3870–3879 (2024)
18. Ke, Z., Sun, J., Li, K., Yan, Q., Lau, R.W.: Modnet: Real-time trimap-free portrait matting via objective decomposition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 1140–1147 (2022)
19. Li, J., Goel, V., Ohanyan, M., Navasardyan, S., Wei, Y., Shi, H.: Vmformer: End-to-end video matting with transformer. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 6678–6687 (2024)
20. Li, J., Jain, J., Shi, H.: Matting anything. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1775–1785 (2024)
21. Li, J., Ohanyan, M., Goel, V., Navasardyan, S., Wei, Y., Shi, H.: Videomatt: A simple baseline for accessible real-time video matting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2177–2186 (2023)
22. Li, J., Ma, S., Zhang, J., Tao, D.: Privacy-preserving portrait matting. In: Proceedings of the 29th ACM international conference on multimedia. pp. 3501–3509 (2021)
23. Li, J., Zhang, J., Maybank, S.J., Tao, D.: Bridging composite and real: towards end-to-end deep image matting. International Journal of Computer Vision **130**(2), 246–266 (2022)
24. Li, J., Zhang, J., Tao, D.: Deep automatic natural image matting. arXiv preprint arXiv:2107.07235 (2021)
25. Li, J., Zhang, J., Tao, D.: Referring image matting. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22448–22457 (2023)
26. Li, X., Yang, Z., Quan, R., Yang, Y.: Drip: Unleashing diffusion priors for joint foreground and alpha prediction in image matting. Advances in Neural Information Processing Systems **37**, 79868–79888 (2024)
27. Lin, S., Ryabtsev, A., Sengupta, S., Curless, B.L., Seitz, S.M., Kemelmacher-Shlizerman, I.: Real-time high-resolution background matting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8762–8771 (2021)
28. Lin, S., Yang, L., Saleemi, I., Sengupta, S.: Robust high-resolution video matting with temporal guidance. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 238–247 (2022)
29. Liu, C., Ding, H., Jiang, X.: Towards enhancing fine-grained details for image matting. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 385–393 (January 2021)
30. Liu, C., Ding, H., Jiang, X.: GRES: Generalized referring expression segmentation. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR). pp. 23592–23601 (2023)
31. Liu, C., Jiang, X., Ding, H.: Primitivenet: decomposing the global constraints for referring segmentation. Visual Intelligence **2**(1), 16 (2024)

32. Liu, Q., Zhang, S., Meng, Q., Zhong, B., Liu, P., Yao, H.: End-to-end human instance matting. IEEE Transactions on Circuits and Systems for Video Technology **34**(4), 2633–2647 (2023)
33. Liu, R.: Enhancing image matting in real-world scenes with mask-guided iterative refinement. arXiv preprint arXiv:2502.17093 (2025)
34. Liu, Y., Xie, J., Shi, X., Qiao, Y., Huang, Y., Tang, Y., Yang, X.: Tripartite information mining and integration for image matting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7555–7564 (2021)
35. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
36. Park, G., Son, S., Yoo, J., Kim, S., Kwak, N.: Matteformer: Transformer-based image matting via prior-tokens. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11696–11706 (2022)
37. Park, K., Woo, S., Oh, S.W., Kweon, I.S., Lee, J.Y.: Mask-guided matting in the wild. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1992–2001 (2023)
38. Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 724–732 (2016)
39. Qiao, Y., Liu, Y., Yang, X., Zhou, D., Xu, M., Zhang, Q., Wei, X.: Attention-guided hierarchical structure aggregation for image matting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13676–13685 (2020)
40. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. In: International Conference on Learning Representations. vol. 2025, pp. 28085–28128 (2025)
41. Seong, H., Oh, S.W., Price, B., Kim, E., Lee, J.Y.: One-trimap video matting. In: European Conference on Computer Vision. pp. 430–448. Springer (2022)
42. Sharma, R., Deora, R., Vishvakarma, A.: Alphanet: An attention guided deep network for automatic image matting. In: 2020 International Conference on Omni-layer Intelligent Systems (COINS). pp. 1–8. IEEE (2020)
43. Sun, Y., Tang, C.K., Tai, Y.W.: Human instance matting via mutual guidance and multi-instance refinement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2647–2656 (2022)
44. Sun, Y., Tang, C.K., Tai, Y.W.: Ultrahigh resolution image/video matting with spatio-temporal sparsity. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14112–14121 (2023)
45. Sun, Y., Wang, G., Gu, Q., Tang, C.K., Tai, Y.W.: Deep video matting via spatio-temporal alignment and aggregation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6975–6984 (2021)
46. Tan, G., Chen, H., Qi, J.: A novel image matting method using sparse manual clicks. Multimedia Tools and Applications **75**(17), 10213–10225 (2016)
47. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: European conference on computer vision. pp. 402–419. Springer (2020)
48. Wang, Y., Xu, B., Li, Z., Huang, H., Lu, C., Guo, Y.: Video object matting via hierarchical space-time semantic guidance. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5120–5129 (2023)

49. Wei, T., Chen, D., Zhou, W., Liao, J., Zhao, H., Zhang, W., Yu, N.: Improved image matting via real-time user clicks and uncertainty estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15374–15383 (2021)
50. Xu, N., Price, B., Cohen, S., Huang, T.: Deep image matting. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2970–2979 (2017)
51. Yang, D., Wang, B., Li, W., Lin, Y., He, C.: Exploring the interactive guidance for unified and effective image matting. ACM Transactions on Multimedia Computing, Communications and Applications **22**(2), 1–24 (2026)
52. Yang, P., Zhou, S., Zhao, J., Tao, Q., Loy, C.C.: Matanyone: Stable video matting with consistent memory propagation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 7299–7308 (2025)
53. Yang, X., Qiao, Y., Chen, S., He, S., Yin, B., Zhang, Q., Wei, X., Lau, R.W.: Smart scribbles for image matting. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) **16**(4), 1–21 (2020)
54. Yao, J., Wang, X., Ye, L., Liu, W.: Matte anything: Interactive natural image matting with segment anything model. Image and Vision Computing **147**, 105067 (2024)
55. Yu, H., Xu, N., Huang, Z., Zhou, Y., Shi, H.: High-resolution deep image matting. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 3217–3224 (2021)
56. Yu, Q., Zhang, J., Zhang, H., Wang, Y., Lin, Z., Xu, N., Bai, Y., Yuille, A.: Mask guided matting via progressive refinement network. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1154–1163 (2021)
57. Yu, Z., Li, X., Huang, H., Zheng, W., Chen, L.: Cascade image matting with deformable graph refinement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7167–7176 (2021)
58. Zhang, C., Hu, Y., Ding, H., Shi, H., Zhao, Y., Wei, Y.: Learning trimaps via clicks for image matting. IEEE Transactions on Multimedia **27**, 8668–8678 (2025). https://doi.org/10.1109/TMM.2025.3607710
59. Zhang, H., Wu, D., Shao, Y., Sang, N., Gao, C.: Object-aware video matting with cross-frame guidance. arXiv preprint arXiv:2503.01262 (2025)
60. Zhang, Y., Gong, L., Fan, L., Ren, P., Huang, Q., Bao, H., Xu, W.: A late fusion cnn for digital matting. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7469–7478 (2019)
61. Zhang, Y., Wang, C., Cui, M., Ren, P., Xie, X., Hua, X.S., Bao, H., Huang, Q., Xu, W.: Attention-guided temporally coherent video object matting. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 5128–5137 (2021)
62. Zhou, Y., Zhou, L., Lam, T.L., Xu, Y.: Semantic-guided automatic natural image matting with trimap generation network and light-weight non-local attention. arXiv preprint arXiv:2103.17020 (2021)