VideoMaMa: Mask-Guided Video Matting via Generative Prior

Supplementary Material

Overview

This supplementary material provides additional details on implementation, evaluation metrics, and qualitative results that could not be included in the main paper due to space constraints.

We strongly recommend viewing the attached interactive HTML viewer. As static PDF figures cannot adequately demonstrate temporal consistency or fine-grained motion details, reviewing the HTML results is pivotal for a comprehensive assessment of our method's performance.

The document is organized as follows:

- **Section A** introduces the **interactive HTML viewer**. This page facilitates a side-by-side comparison between SA-V and MA-V, and showcases VideoMaMa's performance on complex in-the-wild videos. We **strongly recommend** readers to prioritize this section to observe the superiority of our approach.
- **Section B** elaborates on the training configurations for SAM2-Matte and details the architectural design of the VideoMaMa feature injection module.
- **Section C** provides the formal definition and algorithm for the Trimap-based Mean Absolute Difference (MAD-T) metric, designed to evaluate transition region accuracy frame-by-frame.
- **Section D** presents a comparative analysis between SAM2-Matte and the original binary SAM2, empirically demonstrating the necessity and effectiveness of the MA-V dataset.
- **Section G** analyzes the limitations of VideoMaMa, specifically discussing failure cases arising from reliance on semantically inaccurate instance masks.

A. Additional Qualitative Results

Due to the temporal nature of video matting, static PDF figures cannot fully capture the stability and coherence of the matte. To facilitate a comprehensive evaluation of our method, we provide an interactive HTML viewer in the attached supplementary material. This viewer allows for a direct, frame-by-frame comparison between our proposed VideoMaMa, and the comparison between SA-V [31] and MA-V.

We refer reviewers to **`main.html`** included in the supplementary zip file. This interactive page presents:

- Side-by-side video comparisons demonstrating MA-V quality based on ground truth SA-V mask.
- Additional qualitative results of VideoMaMa on in-the-wild videos that include complex motion, intrinsic boundaries and occlusion, which highlights the robustness of our VideoMaMa.

Please open `main.html` in any standard web browser to view these results.

B. Additional Implementation Details

VideoMaMa To obtain in-the-wild results from VideoMaMa, we require input masks for all frames in the video sequence. We employ a point prompt on the first frame, which is then propagated throughout the entire video using SAM2's tracking capability.

For quantitative evaluation on benchmark datasets, we binarize the ground-truth mask from the first frame to serve as the initial prompt for SAM2, subsequently propagating it across all remaining frames to generate the input mask sequence for VideoMaMa.

In the VideoMaMa architecture, we extract semantic features using DINOv2 [39] and inject them into the first upsampling block (Upblock 1) of the Stable Video Diffusion (SVD) [1] backbone. The feature injection is performed through a two-layer MLP projection module that aligns the DINOv3 feature dimensions with the SVD decoder architecture.

SAM2-Matte We implement SAM2-Matte based on the SAM2 [31] Base-Plus architecture, initialized with official pretrained weights. The model is trained for 100,000 iterations with a batch size of 4, employing a mixed training strategy that combines both video and image samples. The composition of existing datasets utilized in this training phase is detailed in Table 7. For optimization, we adopt a combination of L1 and Laplacian pyramid losses. The base learning rate is set to 5×10^{-6}, while a separate learning rate of 3×10^{-6} is applied specifically to the image encoder. Training is conducted with a cosine learning rate scheduler.

Table 7. **Datasets used to train VideoMaMa and SAM2-Matte.**

Category	Datasets
Image	DVM, AM-2k, DAViD, P3M-10k, RefMatte
Video	CRGNN, VideoMatte240K, DVM
Background	VideoMatting108, DVM

C. Trimap-based Mean Absolute Difference

Evaluating matte accuracy in transition regions is critical for video matting, particularly for capturing fine details like hair or motion blur. Standard metrics often dilute these errors by averaging over large background regions. To address this, we evaluate MAD-T, which restricts the Mean Absolute Difference calculation specifically to the unknown

Algorithm 1 Frame-wise Calculation of MAD-T

Require: Pred. $\hat{\alpha}_t$, GT $\alpha_{gt,t}$, Kernel $k = 10$
Ensure: MAD-T Score $\mathcal{S}_t$ for frame t

Initialize structuring element $K \leftarrow \text{Ellipse}(k, k)$
▷ Define Certain Regions via Erosion
$\mathcal{M}_{fg} \leftarrow \mathbb{I}(\alpha_{gt,t} = 255)$ ▷ Binary Foreground
$\mathcal{M}_{bg} \leftarrow \mathbb{I}(\alpha_{gt,t} = 0)$ ▷ Binary Background
$\mathcal{C}_{fg} \leftarrow \mathcal{M}_{fg} \ominus K$ ▷ Erode to ensure certainty
$\mathcal{C}_{bg} \leftarrow \mathcal{M}_{bg} \ominus K$ ▷ Erode to ensure certainty
▷ Isolate Unknown Region (Trimap)
$\mathcal{U}_t \leftarrow \neg(\mathcal{C}_{fg} \cup \mathcal{C}_{bg})$
$N_t \leftarrow \sum \mathcal{U}_t$ ▷ Pixel count in unknown region
▷ Compute Error in Unknown Region
if $N_t > 0$ **then**
  $\mathcal{D}_t \leftarrow |\hat{\alpha}_t - \alpha_{gt,t}|$
  $\mathcal{S}_t \leftarrow \frac{1}{N_t} \sum_{(x,y)\in\mathcal{U}_t} \mathcal{D}_{t,x,y} \times 1000$
else
  $\mathcal{S}_t \leftarrow 0$
end if
return $\mathcal{S}_t$

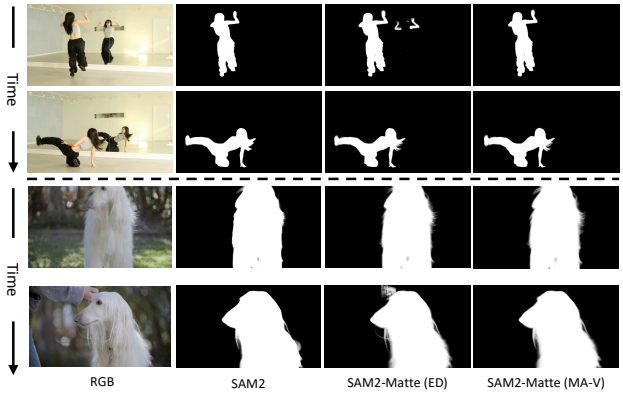

Figure 7. **Comparison on a real video between SAM2-Matte and SAM2** [31] with sigmoid function on mask logit to simulate alpha matte generation. ED denotes that existing dataset.

region. The complete calculation procedure is detailed in Algorithm 1. As ground-truth trimaps are rarely available for video datasets, we generate pseudo-trimaps dynamically for each frame t. We derive binary foreground and background masks from the ground-truth alpha $\alpha_{gt}^{(t)}$ and apply morphological erosion using a 10×10 elliptical structuring element. The unknown region $\mathcal{U}^{(t)}$ is defined as the set of pixels excluded from both eroded masks. The final MAD-T score is reported as the temporal average of the per-frame error within $\mathcal{U}^{(t)}$, scaled by 10^3.

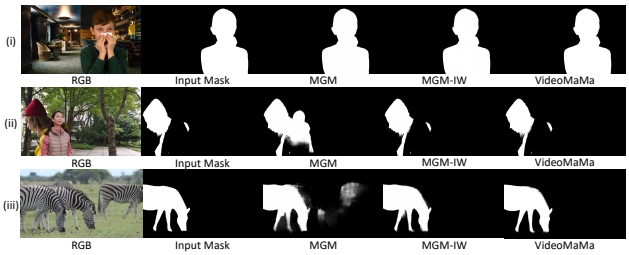

Figure 8. Comparison with MGM failure cases. (i) YouTubeMatte (ii) V-HiM60 (iii) In-the-wild sample.

D. Comparison with SAM2

SAM2 [31] generates binary masks by computing the dot product between image features and prompt tokens that undergo cross-attention in SAM's decoder blocks, followed by thresholding at a specified value. By removing the thresholding operation and retaining only the sigmoid activation, the model can simulate alpha matte generation by mapping the mask logits to the continuous range [0,1], which forms the basis of our SAM2-Matte approach. We compare our SAM2-Matte with the original SAM2 in Figure 7 to demonstrate the effectiveness of the MA-V dataset, illustrating that without fine-tuning on MA-V, SAM2 cannot generate robust video matting results.

E. MGM failure cases

The high errors for MGM in Table 2 arise because it is tuned for single-instance matting, leading to instability on multi-instance video benchmarks, even when the input is already reasonably close (Fig. 8).

F. Efficiency

Table 8. GFLOPs comparison. **(s)**: per step, **(T)**: Total, **(/F)**: per Frame.

Model	MaGGIe	SVD	SAM 2	MatAny-12	MatAny-22
GFLOPs	560.3 **(s)**	30,983 **(s)**	6,403**(T)**/534**(/F)**	2,632**(T)**/219**(/F)**	4,825**(T)**/219**(/F)**

We calculated GFLOPs, throughput (runtime), and memory usage in Table 8. All values are computed on a 12-frame video clip at a spatial resolution of 576×1024. Experiment conducted on RTX A6000.

G. Limitation

VideoMaMa While VideoMaMa can generate high-quality video mattes from rough or imperfect masks, it cannot generate mattes for regions where completely wrong mask is provided (*e.g.* capturing wrong instance). As shown in Figure 9, when the input mask incorrectly captures the wrong instance, VideoMaMa propagates this error to the output. Our method successfully refines masks that lack fine-grained details but struggles when the input mask is

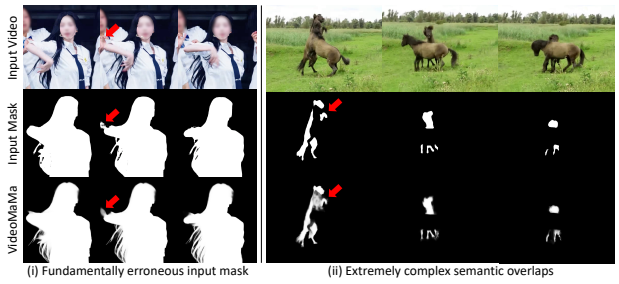

Figure 9. **Limitations.** As highlighted by the red arrows, the model struggles to refine the matte when the input guidance mask exhibits significant errors from the input mask.

fundamentally incorrect, such as when it captures an entirely different object instance. This limitation is inherent to mask-guided approaches, as the model relies on the mask to define the target foreground object.

SAM2-Matte Conducting alpha matte generation in high-resolution is essential for getting high quality alpha matte. However, the standard SAM2 architecture has a structural limitation for this task due to its mask decoder design. Specifically, SAM2 generates segmentation masks at a resolution of 64×64, which is subsequently upsampled to match the input dimensions. This resolution is substantially lower than those employed by other matting methods [17, 49], which typically predict alpha mattes at significantly higher resolutions to preserve fine-grained details such as hair strands and object boundaries. Consequently, directly applying SAM2 for video matting results in the loss of critical high-frequency information necessary for accurate alpha matte estimation. This architectural constraint poses a fundamental challenge for adapting SAM2 to the video matting task, where precise boundary delineation and detail preservation are paramount in getting alpha value of [0,1].