Image Matting without Matting-Specific Annotations via Eikonal Fields

Tao Huang[1], Yi Wang[2], Xiaohong Zhang[1]*, Wei Zhang[1], Tao Yin[1]
Zhibin Zhang[1], Qiuli Wang[3]
[1]School of Big Data and Software Engineering,
Chongqing University
[2]College of Mathematics and Statistics,
Chongqing University
[3]Department of Radiology,
The First Affiliated Hospital of Army Medical University
{taohuang,yiwang,yintao}@stu.cqu.edu.cn,
{xhongz, weizhang0823, zbinz}@cqu.edu.cn, wangqiuli@tmmu.edu.cn

Abstract

Image matting aims to estimate a continuous alpha matte that separates foreground objects from the background. However, existing deep matting methods still rely heavily on task-specific supervision for matting, including alpha mattes and trimaps, which are costly to annotate and limit their applicability in annotation-scarce scenarios. To address this limitation, we propose FreeMatte, a novel framework that learns to predict high-quality alpha mattes without relying on alpha matte or trimap supervision. The proposed framework formulates the construction of matting supervision as a foreground-driven wavefront propagation process. Starting from a coarse localization prior generated by an off-the-shelf segmentation foundation model, it solves an Eikonal equation constructed from the prior and the input image itself to obtain an arrival-time field that encodes how readily each pixel can be reached from the coarse foreground source. The arrival-time field is then used to identify high-confidence background pixels for sparse supervision and to modulate local consistency constraints, providing indirect supervision for alpha prediction. Experiments on AM-2K, AIM-500, P3M-500-NP, and RefMatte-RW-100 demonstrate the effectiveness of FreeMatte, which achieves competitive performance on AM-2K and P3M-500-NP.

1 Introduction

Image matting [1] is a fundamental task in computer vision that aims to softly separate foregrounds from backgrounds by estimating a continuous opacity map, known as the alpha matte [2]. Since an input image alone cannot uniquely solve for the foreground, background, and opacity, estimating the alpha matte from a single image is an ill-posed problem.

To make the problem tractable, matting methods typically introduce additional constraints or priors to reduce the solution space. Traditional methods impose these constraints through user-provided guidance, such as trimaps [3] or scribbles [4], together with hand-crafted assumptions [3, 5]. Although traditional methods can achieve good results in controlled settings, they often lack flexibility due to their reliance on manual guidance and fixed assumptions, and may struggle in complex scenes. Recent deep learning methods instead learn data-driven priors from matting-specific supervision, such as alpha mattes [6, 7] or trimaps [8], and have substantially improved matting quality. They have also

*Corresponding author: xhongz@cqu.edu.cn.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

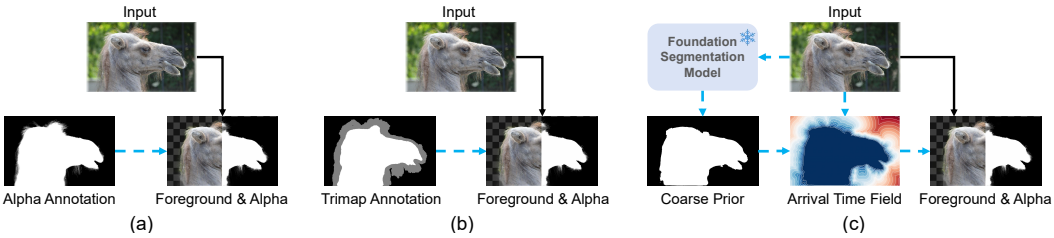

Figure 1: Comparison of different supervision paradigms in image matting. Existing deep matting methods rely on matting-specific supervision during training, such as (a) dense alpha mattes or (b) trimaps. In contrast, (c) FreeMatte constructs supervision from input images and coarse localization priors generated by a pretrained segmentation model, without requiring alpha mattes or trimaps as training targets or loss supervision.

made image matting more flexible in practice, extending the guidance form beyond traditional trimaps to masks [9], point prompts, and box prompts [10], while further enabling automatic matting [11]. As a result, deep learning has become the prevailing paradigm in modern image matting.

However, despite this progress, the success of modern deep matting models still relies heavily on dense, matting-specific supervision. In particular, supervised methods typically require alpha mattes or trimaps (Fig. 1(a) and (b)) for training, which are costly to annotate at scale. This dependence limits the scalability and applicability of image matting when matting-specific annotations are difficult to obtain. This raises a natural question: ***Does high-quality deep image matting necessarily require dense matting-specific annotations?***

To answer this question, we revisit the supervision required by the matting problem itself. Existing deep matting methods often formulate matting as a dense regression task that directly predicts the alpha matte. We argue that this formulation couples two distinct aspects of the problem: foreground localization and continuous transparency inference. The latter is the truly matting-specific component. In contrast, the former is primarily a semantic localization problem, for which modern segmentation foundation models [12, 13] can already provide coarse localization priors.

In this paper, we introduce FreeMatte (Fig. 1(c)), an image matting framework based on Eikonal fields that learns without matting-specific supervision [14]. The key insight is that, given coarse foreground localization, the supervision structure needed for continuous transparency inference can be recovered from the coarse prior and the input image itself, rather than being provided by dense alpha mattes or trimaps. To this end, FreeMatte casts supervision construction as a foreground-driven wavefront propagation process. The coarse prior initializes the propagation source, and the resistance function combines prior-induced spatial constraints with image-derived color and structural cues to control wavefront propagation. Solving the resulting Eikonal equation [15] produces an arrival-time field that serves as a content-aware bridge between coarse localization and alpha-oriented supervision. Pixels are then constrained according to their propagation from the coarse foreground cue, allowing the model to learn from field-induced supervision, thereby eliminating the need for alpha mattes or trimaps during training. Overall, our main contributions are summarized as follows:

- To the best of our knowledge, FreeMatte is the first promptable image matting framework trained without matting-specific supervision, using input images and coarse localization priors from a frozen pretrained segmentation model.
- We propose FreeMatte, an Eikonal-field-based framework that converts coarse localization priors and image cues into confidence-aware sparse supervision for learning alpha mattes without matting-specific annotations.
- We evaluate FreeMatte on AM-2K, AIM-500, P3M-500-NP, and RefMatte-RW-100. The results show that FreeMatte achieves competitive performance with fully supervised methods on AM-2K and P3M-500-NP without alpha matte or trimap supervision.

2 Related Work

Segmentation Models. Recent segmentation models [16–19] enable accurate mask generation across diverse scenes and settings, providing semantic region priors for downstream vision tasks. Among them, the Segment Anything Model (SAM) family [12, 13, 20] has emerged as a representative class of foundation models due to its strong generalization and flexible prompting capability. Although such models provide only coarse semantic regions, they still provide informative cues about object location and approximate boundaries, making them useful priors for image matting.

Image Matting. In contrast to image segmentation, image matting estimates a dense alpha matte with continuous opacity transitions and fine boundary details. Traditional image matting methods [5, 4, 21, 3, 22, 23] typically estimate alpha with additional constraints, such as color sampling, local smoothness, affinity propagation, or Laplacian regularization. However, most rely on a trimap or similar user-provided guidance to define known and unknown regions, which limits their flexibility and often degrades performance in challenging scenes. Deep image matting methods [24, 7, 6, 25–29, 9, 30, 11, 31, 10, 32–37, 8] replace hand-crafted pipelines with data-driven models and have substantially improved matting quality in complex scenes. Their guidance forms have also become more flexible, extending from trimaps [7, 6, 26] to masks [9, 30] and interactive prompts [27, 11, 10, 33]. Recent image matting methods built on foundation models further broaden this design space, including SAM-based methods [27, 10, 34], ViT-based methods [11, 32], and diffusion-based methods [28, 33, 31]. Despite these advancements, most existing deep matting methods are still trained with costly alpha matte annotations. VMFM [38] explores self-supervised matting for human-object interaction, but its foreground prediction networks are pretrained with ground-truth alpha mattes. A notable recent exception is DDC-Matting [8], which solely relies on trimap annotations and color consistency constraints for training. Motivated by the complementary strengths of segmentation foundation models in localization and propagation-based methods in exploiting image structures, we explore a new paradigm that replaces matting-specific annotations with supervision constructed from coarse localization priors and image cues.

Eikonal-based Methods in Vision. Eikonal fields have been widely used in computer vision. Classical approaches typically solve the Eikonal equation directly using level set methods or fast marching algorithms, and apply the resulting arrival-time or geodesic distance fields to tasks such as geodesic active contours, interactive image segmentation, and boundary-aware matting pipelines [14, 39, 40, 4]. More recently, learning-based methods have incorporated the Eikonal equation as a structural constraint in neural implicit representations, novel view synthesis, and travel-time prediction, where the goal remains to learn or approximate the underlying field satisfying the Eikonal property [41–45]. Different from these works, we utilize the Eikonal field not as the prediction target, but as a unified propagation space to integrate constraints required for image matting.

3 Method

We introduce FreeMatte, an image matting framework that learns without matting-specific supervision. As shown in Fig. 2, our method constructs sparse supervisory signals by defining an image- and prior-dependent resistance function, solving the corresponding Eikonal equation to obtain an arrival-time field, and converting this field into probabilistic supervision. We present the problem formulation and analysis in Sec. 3.1, introduce the Eikonal formulation in Sec. 3.2, detail the resistance function design in Sec. 3.3, describe the sparse supervision construction in Sec. 3.4, and introduce the training objective in Sec. 3.5.

3.1 Problem Formulation and Analysis

Problem Formulation. Let $I \in \mathbb{R}^{H\times W\times 3}$ denote an input image defined on the image domain Ω, and let $\alpha \in [0,1]^{H\times W}$ denote the target alpha matte. In our setting, matting-specific annotations, such as alpha mattes or trimaps, are not available. Instead, each image is accompanied by a coarse localization prior $S \in \{0,1\}^{H\times W}$, generated by a pretrained segmentation model (e.g., SAM), which contains no opacity information. We then define the coarse foreground support induced by S as $\Omega_f = \{x \in \Omega \mid S(x) = 1\}$, where x denotes a pixel coordinate in Ω. Given training pairs

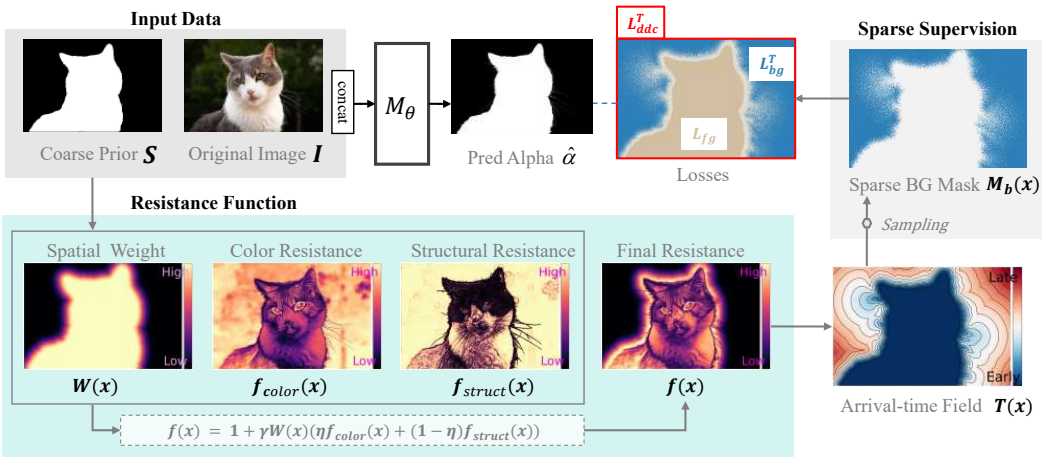

Figure 2: Overview of FreeMatte. Given an image I and a coarse localization prior S, FreeMatte predicts an alpha matte $\hat{\alpha}$ and constructs Eikonal-based supervision in parallel. A final resistance $f(x)$ is built from the spatial weight $W(x)$, color resistance $f_{\text{color}}(x)$, and structural resistance $f_{\text{struct}}(x)$, and is used to compute an arrival-time field $T(x)$. The arrival-time field is converted into a sparse background mask $M_b(x)$ through sampling. The predicted alpha matte is supervised by $\mathcal{L}_{\text{fg}}$, $\mathcal{L}^T_{\text{bg}}$, and $\mathcal{L}^T_{\text{ddc}}$, without using alpha mattes or trimaps.

$\{(I, S)\}$, our goal is to learn a matting model M_θ that predicts a spatially coherent alpha matte $\hat{\alpha} = M_\theta(I, S) \in [0, 1]^{H \times W}$.

Analysis. Formally, trimap-guided matting models regularize the ill-posed matting problem by leveraging explicit spatial partitions: known foreground and background regions provide reliable endpoint constraints, while opacity inference is mainly performed in unknown regions. A recent important development following this paradigm is DDC-Matting [8], which trains matting models without alpha annotations by combining trimap-provided known-region supervision with a directional distance consistency term.

To expose the underlying supervision structure, we rewrite the DDC-Matting objective in the following structural form:

$$\mathcal{L}^{\text{trimap}}_{\text{matting}} = \mathcal{L}^{\text{fg}}_{\text{anchor}} + \mathcal{L}^{\text{bg}}_{\text{anchor}} + \lambda \mathcal{L}_{\text{consistency}}. \tag{1}$$

Here, $\lambda > 0$ is a generic balancing coefficient used in this structural decomposition. $\mathcal{L}^{\text{fg}}_{\text{anchor}}$ and $\mathcal{L}^{\text{bg}}_{\text{anchor}}$ are L1 anchor losses that enforce $\alpha = 1$ on known foreground pixels and $\alpha = 0$ on known background pixels provided by the trimap, respectively. The term $\mathcal{L}_{\text{consistency}}$ denotes a generic consistency constraint that regularizes alpha variations in unknown regions. In DDC-Matting, this consistency term is instantiated as the directional distance consistency loss, denoted by $\mathcal{L}_{\text{ddc}}$ [8].

Eq. (1) reveals an anchor-and-consistency structure behind trimap-guided supervision: foreground and background anchors constrain the two extrema of the alpha solution, while the consistency term regularizes alpha values in unknown regions under these endpoint constraints. However, this structure is not directly provided in our setting, where no trimap annotations are available for loss supervision.

3.2 Eikonal Formulation

To recover the supervision structure identified in Sec. 3.1, we formulate supervision construction as a foreground-driven wavefront propagation process. Under this formulation, the coarse foreground support Ω_f serves as the source region, while the resistance function $f(x)$ specifies the local propagation cost. This leads to an Eikonal equation whose solution is an arrival-time field $T(x)$:

$$\begin{cases} \|\nabla T(x)\| = f(x), & \forall x \in \Omega \setminus \Omega_f, \\ T(x) = 0, & \forall x \in \Omega_f, \end{cases} \tag{2}$$

where Ω denotes the image domain.

Under this formulation, $T(x)$ can be interpreted as the minimum accumulated propagation cost from the coarse foreground source Ω_f to pixel x under the resistance function $f(x)$. Specifically, a pixel obtains a small arrival time if it can be connected to Ω_f through a low-resistance path, and obtains a large arrival time if reaching it requires crossing high-resistance regions.

3.3 Resistance Function Design

The Eikonal formulation provides the propagation framework, while the resistance function $f(x)$ specifies how difficult it is for the wavefront to propagate at each pixel. In our setting, $f(x)$ should encourage propagation through pixels that are compatible with the coarse foreground prior, while assigning larger costs to pixels that are less likely to belong to the foreground. We therefore define $f(x)$ by integrating color, structural, and spatial cues derived from the input image and the coarse foreground prior.

The resistance function is defined as:

$$f(x) = 1 + \gamma W(x)\Big(\eta\, f_{\mathrm{color}}(x) + (1-\eta)\, f_{\mathrm{struct}}(x)\Big), \tag{3}$$

where $f_{\mathrm{color}}(x)$ and $f_{\mathrm{struct}}(x)$ denote the image-dependent color and structural resistance terms. The spatial resistance is composed of the constant base cost 1 and the distance-aware weight $W(x)$. The constant term provides a uniform propagation cost across pixels, while $W(x)$ localizes the image-dependent resistance around the coarse foreground prior. The coefficient γ controls the overall resistance strength, and $\eta \in [0,1]$ balances the color and structural terms.

Color resistance. The color resistance measures the appearance compatibility between each pixel and the coarse foreground support. Let μ_{fg} and Σ_{fg} denote the RGB mean and diagonal covariance estimated over Ω_f, where $\mu_{fg} = \frac{1}{|\Omega_f|}\sum_{x\in\Omega_f} I(x)$ and $\Sigma_{fg} = \mathrm{diag}(\sigma_{fg}^2) + \varepsilon I_3$, where $\sigma_{fg}^2 \in \mathbb{R}^3$ denotes the channel-wise RGB variance computed over the coarse foreground support Ω_f, I_3 is the 3×3 identity matrix, and ε is a small constant for numerical stability. We define the foreground-normalized color deviation and color resistance as:

$$\Delta_{\mathrm{color}}(x) = \left\| \Sigma_{fg}^{-\frac{1}{2}} \big(I(x) - \mu_{fg}\big) \right\|_2, \quad f_{\mathrm{color}}(x) = \frac{\Delta_{\mathrm{color}}(x)}{\max_{y\in\Omega} \Delta_{\mathrm{color}}(y) + \varepsilon}. \tag{4}$$

Here, $\Delta_{\mathrm{color}}(x)$ corresponds to a diagonal Mahalanobis distance, which normalizes RGB deviations by the channel-wise foreground variance. This term assigns larger resistance to pixels that deviate from the coarse foreground prototype, thereby suppressing wavefront leakage into visually incompatible background regions.

Structural resistance. The structural resistance is designed to preserve sensitivity to fine details such as hair strands, thin structures, and object boundaries. We first compute the gradient magnitude $G(x)$ from the grayscale image and normalize it to obtain $G_{\mathrm{norm}}(x)$. We use an adaptive gradient scale $\sigma_g = \frac{1}{|\Omega|}\sum_{y\in\Omega} G_{\mathrm{norm}}(y)$, and define the structural term as:

$$f_{\mathrm{struct}}(x) = \exp\left(-\frac{G_{\mathrm{norm}}(x)^2}{\sigma_g^2 + \varepsilon}\right). \tag{5}$$

Under this design, high-gradient regions induce lower additional resistance. This helps the arrival-time field remain sensitive to fine transition structures, while smooth regions receive higher structural resistance.

Spatial resistance. Spatial resistance is implemented through the constant term 1 in the final resistance function and a distance-aware modulation term $W(x)$. The constant term in Eq. (3) ensures a nonzero propagation cost over the entire image domain, while $W(x)$ makes the color and structural resistance terms more sensitive near the coarse foreground prior and gradually weakens their effect with increasing distance. Let $D(x) = \mathrm{dist}(x, \Omega_f)$ denote the Euclidean distance from pixel x to the coarse foreground support. We define:

$$W(x) = \exp\left(-\frac{D(x)^2}{2(rD_{\max})^2 + \varepsilon}\right), \tag{6}$$

where $D_{\max}$ is the maximum value of D over the image domain, and r controls the effective spatial range of the modulation. This modulation gives larger weights to the color and structural resistance

terms near the foreground prior and suppresses their influence in distant background regions, so that the final resistance gradually falls back to the constant propagation cost.

Given the resistance function defined above, we numerically solve the Eikonal equation in Eq. (2) using the Fast Marching Method (FMM) [15] to obtain the arrival-time field $T(x)$.

3.4 Sparse Supervision

Having computed the arrival-time field $T(x)$, we proceed to reconstruct the missing background anchors identified in Sec. 3.1. Since a larger arrival time inherently reflects higher cumulative resistance from the foreground source, it serves as a reliable indicator of background confidence. We thus transform $T(x)$ into a per-pixel sampling probability for background anchors, from which sparse background supervision is obtained through Bernoulli sampling:

$$p_b(x) = \min\left(\frac{T(x)}{\tau \max_{y\in\Omega} T(y) + \varepsilon}, 1\right), \qquad M_b(x) \sim \text{Bernoulli}(p_b(x)), \tag{7}$$

where τ controls the arrival-time threshold above which pixels are treated as high-confidence background anchors and sampled with probability one.

The mask M_b provides the missing background anchors by enforcing $\hat{\alpha}(x) = 0$ on the selected pixels, i.e., where $M_b(x) = 1$. Together with the foreground prior, it successfully completes the supervision structure required in Eq. (1). An expectation-based interpretation is provided in Appendix B.

3.5 Training Objective

Following the anchor-and-consistency structure identified in Sec. 3.1, we train M_θ with foreground anchoring, sparse background anchoring, and Eikonal-gated transition consistency. We instantiate the foreground anchor term $\mathcal{L}^{\text{fg}}_{\text{anchor}}$ with an L1 loss on Ω_f, yielding $\mathcal{L}_{\text{fg}}$.

For background anchoring, we instantiate the background anchor term $\mathcal{L}^{\text{bg}}_{\text{anchor}}$ using the sparse sampling mask $M_b(x)$ derived from the arrival-time field in Sec. 3.4. The sampled pixels provide sparse background anchors, on which we impose $\hat{\alpha}(x) = 0$:

$$\mathcal{L}^{T}_{\text{bg}} = \frac{1}{\sum_{x\in\Omega} M_b(x) + \varepsilon} \sum_{x\in\Omega} M_b(x)\,|\hat{\alpha}(x)|. \tag{8}$$

Following DDC-Matting, we construct pixel pairs $\mathcal{P} = \{(i, j) \mid j \in \mathcal{N}(i)\}$, where $\mathcal{N}(i)$ contains neighbors selected by RGB similarity within a local window centered at pixel i. I_i and I_j denote the RGB values at pixels i and j, while $\hat{\alpha}_i$ and $\hat{\alpha}_j$ denote their predicted alpha values. We define the Eikonal-gated DDC loss $\mathcal{L}^{T}_{\text{ddc}}$ as:

$$g_i = 1 - \kappa \bar{T}_i, \qquad \mathcal{L}^{T}_{\text{ddc}} = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \left|(\hat{\alpha}_i - \hat{\alpha}_j) - g_i \|I_i - I_j\|_2\right|. \tag{9}$$

Here, $\bar{T}_i \in [0, 1]$ denotes the min-max normalized arrival time at pixel i, and κ controls the attenuation strength. The gate g_i scales the target distance $\|I_i - I_j\|_2$ rather than the loss itself. Larger arrival times reduce this target, encouraging smaller alpha differences in confident background regions and suppressing texture-induced variations.

The overall training objective follows the same structural form as Eq. (1):

$$\mathcal{L}_{\text{matting}} = \mathcal{L}_{\text{fg}} + \mathcal{L}^{T}_{\text{bg}} + \lambda_{\text{c}} \mathcal{L}^{T}_{\text{ddc}}. \tag{10}$$

Here, λ_{c} balances the consistency term. Together, these terms instantiate the anchor-and-consistency supervision structure.

4 Experiment

4.1 Experimental Setup

Implementation Details. We use ViTMatte [6] as the alpha prediction network and SAM-B [12] as a frozen segmentation foundation model to generate coarse localization priors. Unless otherwise

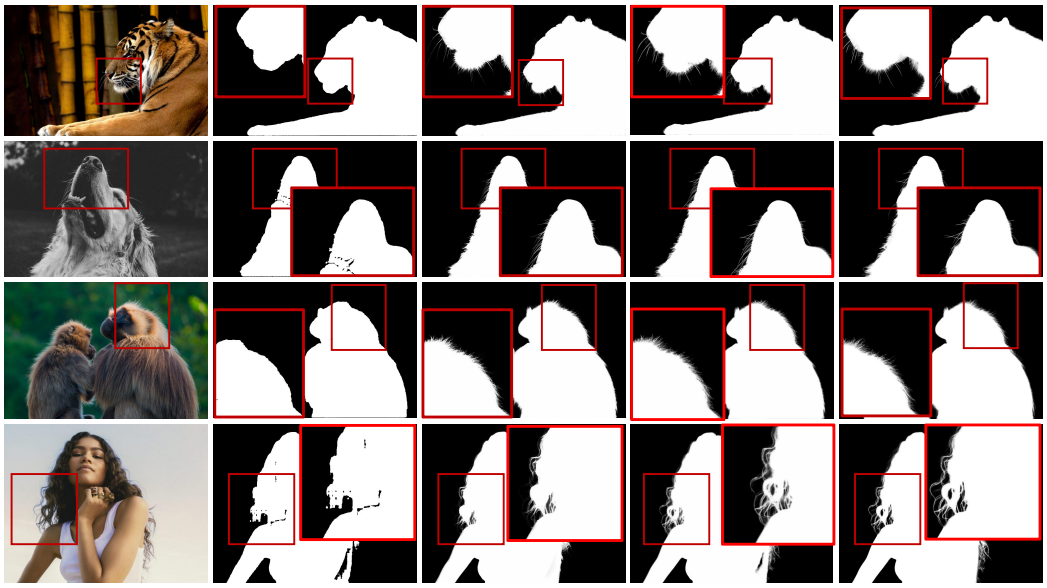

Figure 3: Visual comparison with coarse priors and ground-truth alpha mattes.

Table 1: Main comparison under the point-prompt protocol on four benchmarks. The best and second-best results are marked in bold and underlined, respectively. Results of competing methods are adopted from [33], except for ZIM, whose results are taken from its original paper [32].

Method	Publication	Pretrained Model	Task Annotation	**AM-2K** (animal)					**AIM-500** (natural)				
				MSE ↓	SAD ↓	MAD ↓	Grad ↓	Conn ↓	MSE ↓	SAD ↓	MAD ↓	Grad ↓	Conn ↓
MatAny [27]	IVC'24	SAM-H	Matte	0.0116	32.20	0.0188	15.68	20.39	0.0425	87.05	0.0523	33.44	25.35
MAM [10]	CVPRW'24	SAM-B	Matte	0.0597	141.6	0.0813	22.48	31.52	0.0752	186.5	0.1080	37.48	40.38
SmartMatting [11]	CVPR'24	DINOv2	Matte	0.0302	62.61	0.0366	33.82	15.93	0.0302	66.27	0.0388	46.63	18.77
ZIM [32]	ICCV'25	ViT-B	Matte	0.0226	–	–	15.68	–	0.0424	–	–	27.82	–
LiteSDMatte [33]	ICCV'25	SD2	Matte	0.0095	27.51	0.0161	13.59	17.74	0.0115	34.43	0.0207	**24.32**	19.97
SDMatte [33]	ICCV'25	SD2	Matte	**0.0060**	**17.54**	**0.0104**	**13.17**	**10.86**	**0.0109**	**31.80**	**0.0189**	26.84	**17.51**
FreeMatte	–	SAM-B	Free	0.0089	27.49	0.0160	17.43	11.47	0.0253	58.83	0.0356	35.80	24.58
				P3M-500-NP (human)					**RefMatte-RW-100** (human)				
MatAny [27]	IVC'24	SAM-H	Matte	0.0295	57.33	0.0342	25.95	15.97	0.0118	24.35	0.0137	18.13	4.98
MAM [10]	CVPRW'24	SAM-B	Matte	0.0875	207.5	0.1163	29.43	43.49	0.1651	336.5	0.1896	49.91	27.80
SmartMatting [11]	CVPR'24	DINOv2	Matte	0.0239	50.46	0.0291	28.50	19.64	0.0127	26.75	0.0153	23.01	5.38
LiteSDMatte [33]	ICCV'25	SD2	Matte	**0.0121**	**29.94**	**0.0173**	**16.55**	21.82	0.0096	22.90	0.0131	15.74	7.29
SDMatte [33]	ICCV'25	SD2	Matte	0.0134	32.02	0.0183	20.35	20.76	**0.0091**	**20.45**	**0.0116**	**15.57**	**4.01**
FreeMatte	–	SAM-B	Free	0.0147	34.82	0.0199	20.30	**13.14**	0.0098	31.59	0.0182	22.02	6.71

specified, we adopt point prompts as the default setting, following the same prompt generation method as SDMatte [33] for fair comparison. We set $\lambda_c = 10$, and fix the Eikonal-related hyperparameters as $\gamma = 10.0$, $\eta = 0.8$, $r = 0.1$, $\tau = 0.8$, and $\kappa = 0.5$ across all experiments. Further implementation details are provided in Appendix A.

Datasets and Metrics. Experiments are conducted on four natural image matting benchmarks: AIM-500 [46], AM-2K [47], P3M [48], and RefMatte-RW-100 [49]. They cover diverse natural scenarios, including generic objects, animals, and human portraits. We evaluate matting quality using mean squared error (MSE), sum of absolute differences (SAD), mean absolute difference (MAD), gradient error (Grad) [50], and connectivity error (Conn) [50]. Lower values indicate better performance for all metrics.

4.2 Main Results

Figure 3 presents qualitative results of FreeMatte and the fully supervised ViTMatte baseline under the same point-prompt protocol, alongside the coarse prior masks and ground-truth alpha mattes.

Comparison with Promptable Methods. Table 1 compares FreeMatte with recent promptable matting methods under the point-prompt protocol on four benchmarks. For FreeMatte, a frozen SAM-B converts point prompts into coarse localization priors. The compared baselines use alpha-matte supervision, whereas FreeMatte is optimized without alpha-matte or trimap loss supervision.

FreeMatte achieves competitive results, particularly on AM-2K and P3M-500-NP. On AM-2K, it ranks second among the reported results in MSE, SAD, MAD, and Conn, behind SDMatte. On P3M-500-NP, it achieves the lowest Conn error and the second-lowest Grad error. Compared with SDMatte, FreeMatte obtains a lower Conn error (13.14 vs. 20.76) and a similar Grad error (20.30 vs. 20.35). However, SDMatte achieves lower errors than FreeMatte on all five metrics on AIM-500 and RefMatte-RW-100. These results demonstrate the potential of learning image matting from coarse localization priors and image-derived constraints, while also revealing a remaining gap to stronger fully supervised methods.

Table 2: Matched comparisons with DDC-Matting on AM-2K under two input settings: (a) RGB-only input, where the SAM-B mask is used only to construct training supervision for FreeMatte and is not provided as a network input; and (b) RGB concatenated with the same SAM-B mask.

(a) RGB input

Method	Task Annotation	MSE ↓	SAD ↓	MAD ↓	Grad ↓	Conn ↓
DDC-Matting	Trimap	0.0099	30.26	0.0175	14.98	14.33
DDC-Matting	Free	0.0174	41.40	0.0239	23.26	12.43
FreeMatte	Free	0.0138	35.52	0.0203	17.94	12.31

(b) RGB + mask input

Method	Task Annotation	MSE ↓	SAD ↓	MAD ↓	Grad ↓	Conn ↓
DDC-Matting	Trimap	0.0085	26.71	0.01552	16.38	12.92
DDC-Matting	Free	0.0104	27.66	0.0161	25.50	10.71
FreeMatte	Free	0.0089	27.49	0.0160	17.43	11.47

Comparison with DDC-Matting. A comparison between FreeMatte and DDC-Matting is conducted on AM-2K with RGB input and RGB input concatenated with the SAM-B mask, using the same ViTMatte backbone, training set, and evaluation protocol. Under the Trimap setting, DDC-Matting is trained with a trimap derived from the ground-truth alpha matte, whereas under the Free setting, both DDC-Matting and FreeMatte construct supervision from the same SAM-B prior. As shown in Table 2, FreeMatte outperforms DDC-Matting on all five metrics with RGB input and on four metrics with RGB input concatenated with the SAM-B mask. These results suggest that the performance gains primarily stem from the proposed supervision construction based on the Eikonal field, rather than changes in the prediction architecture.

Comparison with Full Supervision. A comparison between FreeMatte and fully supervised ViTMatte [6] is conducted on AM-2K with the same point prompt setting for a more controlled performance comparison. Both methods use ViTMatte as the prediction backbone, but alpha matte supervision is used to train the fully supervised baseline. As shown in Table 3, fully supervised ViTMatte achieves lower errors across all five metrics, with SAD and Grad of 21.99 and 11.91, respectively, compared with 27.49 and 17.43 for FreeMatte. This comparison quantifies the performance gap between training without matting-specific annotations and full alpha matte supervision.

Table 3: Comparison with full supervision using the same ViTMatte backbone on AM-2K.

Method	Task Annotation	MSE ↓	SAD ↓	MAD ↓	Grad ↓	Conn ↓
ViTMatte	Matte	0.0078	21.99	0.0128	11.91	9.80
FreeMatte	Free	0.0089	27.49	0.0160	17.43	11.47

4.3 Ablation Study

To validate the key design choices of FreeMatte, we conduct ablation studies on AM-2K from two aspects: the contribution of each supervision component in the training objective and the effect of different cues in the Eikonal resistance function.

Overall Ablation. Table 4a validates the contribution of each supervision component. The coarse prior S alone provides object-level localization but lacks accurate boundary modeling, as reflected by its high Grad error of 67.42. Using only $\mathcal{L}_{\mathrm{fg}}$ and $\mathcal{L}_{\mathrm{ddc}}$ for supervision leads to unstable predictions, with an MSE of 0.4396 and an SAD of 921.4. Adding background anchoring through $\mathcal{L}_{\mathrm{bg}}^{\mathrm{base}}$ substantially restores performance when combined with $\mathcal{L}_{\mathrm{fg}}$ and $\mathcal{L}_{\mathrm{ddc}}$, reducing the MSE to 0.0104 and the

Table 4: Main ablations of FreeMatte on AM-2K. (a) Overall ablation. The S-only baseline directly evaluates the binary SAM-B prior as an alpha matte without network training, while all learned variants take (I, S) as input. $\mathcal{L}_{\text{fg}}$ denotes the foreground anchoring loss. $\mathcal{L}_{\text{bg}}^{\text{base}}$ and $\mathcal{L}_{\text{bg}}^{T}$ denote the background anchoring losses whose supports are constructed without and with the Eikonal field, respectively. $\mathcal{L}_{\text{ddc}}^{T}$ denotes the Eikonal-gated DDC loss. (b) Resistance ablation. f_{color}, f_{struct}, and $W(x)$ denote the color resistance, structural resistance, and distance-aware modulation term, respectively.

(a) Overall ablation

Variant	MSE ↓	SAD ↓	MAD ↓	Grad ↓	Conn ↓
S	0.0138	30.50	0.0178	67.42	**8.352**
$S + \mathcal{L}_{\text{fg}} + \mathcal{L}_{\text{bg}}^{\text{base}}$	0.0173	40.35	0.0235	87.56	12.55
$S + \mathcal{L}_{\text{fg}} + \mathcal{L}_{\text{ddc}}$	0.4396	921.4	0.5345	37.47	128.8
$S + \mathcal{L}_{\text{fg}} + \mathcal{L}_{\text{bg}}^{T}$	0.0119	31.92	0.0186	39.99	14.98
$S + \mathcal{L}_{\text{fg}} + \mathcal{L}_{\text{bg}}^{\text{base}} + \mathcal{L}_{\text{ddc}}$	0.0104	27.66	0.0161	25.50	10.71
$S + \mathcal{L}_{\text{fg}} + \mathcal{L}_{\text{bg}}^{T} + \mathcal{L}_{\text{ddc}}$	0.0092	30.31	0.0177	17.60	11.77
$S + \mathcal{L}_{\text{fg}} + \mathcal{L}_{\text{bg}}^{T} + \mathcal{L}_{\text{ddc}}^{T}$	**0.0089**	**27.49**	**0.0160**	**17.43**	11.47

(b) Resistance ablation

f_{color}	f_{struct}	$W(x)$	MSE ↓	SAD ↓	MAD ↓	Grad ↓	Conn ↓
✓			0.0093	31.25	0.0182	17.60	11.85
	✓		0.0092	30.72	0.0179	17.89	11.84
		✓	0.0090	28.92	0.0168	17.58	11.71
✓	✓	✓	**0.0089**	**27.49**	**0.0160**	**17.43**	**11.47**

SAD to 27.66. Replacing $\mathcal{L}_{\text{bg}}^{\text{base}}$ with $\mathcal{L}_{\text{bg}}^{T}$ further reduces the MSE to 0.0092 and the Grad error to 17.60. Finally, with $\mathcal{L}_{\text{bg}}^{T}$ fixed, replacing $\mathcal{L}_{\text{ddc}}$ with $\mathcal{L}_{\text{ddc}}^{T}$ improves all five metrics.

Resistance Function Ablation. Table 4b shows that the color, structural, and spatial components of the resistance function provide complementary information. The color term models foreground appearance compatibility, the structural term preserves boundary-sensitive cues, and the spatial weight $W(x)$ controls the effective propagation range. Combining all three components yields the best performance, providing a more informative arrival-time field for supervision construction.

4.4 Different Priors Analysis

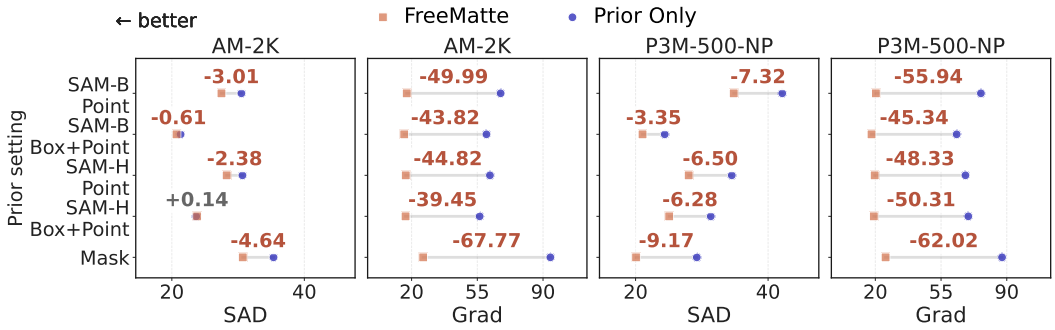

Figure 4: Prior source analysis on AM-2K and P3M-500-NP. For each prior source and prompting strategy, *Prior Only* directly evaluates the raw prior, whereas *FreeMatte* takes the same prior as input and predicts the final alpha matte. The gap between the two endpoints reflects how effectively our framework converts coarse priors into improved matting predictions.

As shown in Fig. 4, FreeMatte consistently reduces Grad across different prior sources, indicating that it preserves coarse localization while substantially improving boundary details. On both AM-2K and P3M-500-NP, FreeMatte largely decreases SAD as well, especially when using Mask priors. For example, on P3M-500-NP, SAD is reduced from 29.24 to 20.07 and Grad from 87.49 to 25.47. These results show that FreeMatte is robust to different prior qualities and can effectively convert coarse priors into accurate boundary-aware alpha mattes.

4.5 Pseudo Trimap Replacement Study

We further compare FreeMatte with pseudo-trimap variants constructed using dilation/erosion radii $\rho \in \{3, 5, 10, 15, 20, 30, 40\}$. Each pseudo-trimap is generated from the same coarse prior and

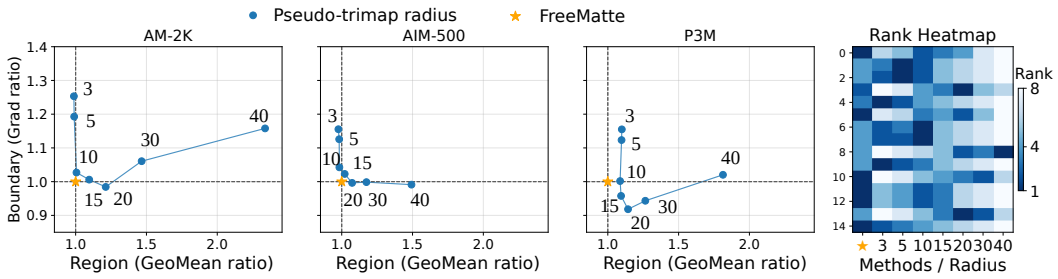

Figure 5: Pseudo-trimap replacement study on AM-2K, AIM-500, and P3M-500-NP. Numbered points denote pseudo-trimap variants with different dilation/erosion radii ρ, and the star denotes FreeMatte. All values are normalized by FreeMatte at $(1, 1)$; lower values are better. The heatmap reports metric-wise rankings across datasets and radii.

replaces our Eikonal-derived sparse background anchoring with pseudo-trimap-based known-region supervision. As shown in Fig. 5, pseudo-trimaps exhibit a clear region-boundary trade-off, and the best radius varies across datasets and metrics. In contrast, FreeMatte achieves a more balanced performance with a single parameter setting, avoiding dataset- or metric-specific radius tuning.

4.6 Hyperparameter Analysis

Fig. 6 analyzes the sensitivity of FreeMatte to four key hyperparameters: γ, η, τ, and r. We vary each hyperparameter individually while keeping the others fixed, and evaluate the resulting performance using SAD and Grad. As shown in Fig. 6, FreeMatte exhibits relatively stable performance across the tested ranges, with only modest variations in both metrics. These results indicate that the proposed supervision framework is not overly sensitive to individual hyperparameter choices.

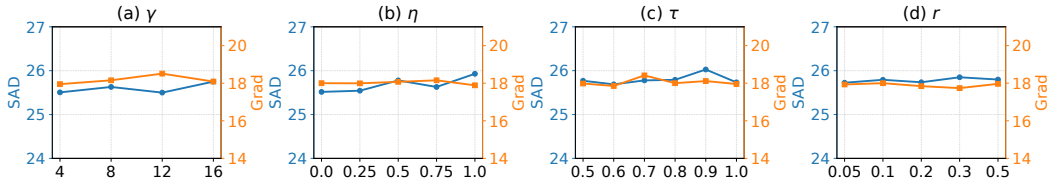

Figure 6: Hyperparameter sensitivity of FreeMatte under representative settings. We vary one hyperparameter at a time and report SAD and Grad for each setting. The four subplots correspond to (a) resistance strength γ, (b) color-structure ratio η, (c) sparse supervision threshold τ, and (d) spatial bandwidth ratio r.

5 Conclusion

We present FreeMatte, an image matting framework that converts coarse localization priors and intrinsic image cues into effective supervision through image-conditioned Eikonal fields. FreeMatte achieves competitive performance on diverse natural image matting benchmarks without relying on alpha mattes or trimaps during training, all while utilizing a single set of hyperparameters. Future work will explore transferring FreeMatte to broader application domains, particularly annotation-scarce fields like medical imaging, where acquiring dense ground truths is costly and difficult.

Acknowledgments and Disclosure of Funding

This work was supported in part by the National Natural Science Foundation of China under Grant 62276033. The experiments in this work were conducted on the Huawei Cloud AI Compute Service. We thank Huawei Cloud for providing reliable computing resources.

References

[1] Mark A Ruzon and Carlo Tomasi. Alpha estimation in natural images. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662)*, volume 1, pages 18–25. IEEE, 2000.

[2] Jizhizi Li, Jing Zhang, and Dacheng Tao. Deep image matting: A comprehensive survey. *arXiv preprint arXiv:2304.04672*, 2023.

[3] Anat Levin, Dani Lischinski, and Yair Weiss. A closed-form solution to natural image matting. *IEEE transactions on pattern analysis and machine intelligence*, 30(2):228–242, 2007.

[4] Xue Bai and Guillermo Sapiro. Geodesic matting: A framework for fast interactive image and video segmentation and matting. *International journal of computer vision*, 82(2):113–132, 2009.

[5] Qifeng Chen, Dingzeyu Li, and Chi-Keung Tang. Knn matting. *IEEE transactions on pattern analysis and machine intelligence*, 35(9):2175–2188, 2013.

[6] Jingfeng Yao, Xinggang Wang, Shusheng Yang, and Baoyuan Wang. Vitmatte: Boosting image matting with pre-trained plain vision transformers. *Information Fusion*, 103:102091, 2024.

[7] GyuTae Park, SungJoon Son, JaeYoung Yoo, SeHo Kim, and Nojun Kwak. Matteformer: Transformer-based image matting via prior-tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11696–11706, 2022.

[8] Wenze Liu, Zixuan Ye, Hao Lu, Zhiguo Cao, and Xiangyu Yue. Training matting models without alpha labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 5604–5612, 2025.

[9] Qihang Yu, Jianming Zhang, He Zhang, Yilin Wang, Zhe Lin, Ning Xu, Yutong Bai, and Alan Yuille. Mask guided matting via progressive refinement network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1154–1163, 2021.

[10] Jiachen Li, Jitesh Jain, and Humphrey Shi. Matting anything. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1775–1785. IEEE, 2024.

[11] Zixuan Ye, Wenze Liu, He Guo, Yujia Liang, Chaoyi Hong, Hao Lu, and Zhiguo Cao. Unifying automatic and interactive matting with pretrained vits. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25585–25594, 2024.

[12] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

[13] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[14] Vicent Caselles, Ron Kimmel, and Guillermo Sapiro. Geodesic active contours. *International journal of computer vision*, 22(1):61–79, 1997.

[15] James A Sethian. A fast marching level set method for monotonically advancing fronts. *proceedings of the National Academy of Sciences*, 93(4):1591–1595, 1996.

[16] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.

[17] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 603–612, 2019.

[18] Jitesh Jain, Jiachen Li, Mang Tik Chiu, Ali Hassani, Nikita Orlov, and Humphrey Shi. Oneformer: One transformer to rule universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2989–2998, 2023.

[19] Dantong Niu, Xudong Wang, Xinyang Han, Long Lian, Roei Herzig, and Trevor Darrell. Unsupervised universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22744–22754, 2024.

[20] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

[21] Leo Grady, Thomas Schiwietz, Shmuel Aharon, and Rüdiger Westermann. Random walks for interactive alpha-matting. In *Proceedings of VIIP*, volume 2005, pages 423–429, 2005.

[22] Jian Sun, Jiaya Jia, Chi-Keung Tang, and Heung-Yeung Shum. Poisson matting. In *ACM SIGGRAPH 2004 Papers*, pages 315–321. 2004.

[23] Yung-Yu Chuang, Brian Curless, David H Salesin, and Richard Szeliski. A bayesian approach to digital matting. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 2, pages II–II. IEEE, 2001.

[24] Ning Xu, Brian Price, Scott Cohen, and Thomas Huang. Deep image matting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2970–2979, 2017.

[25] Jinlin Liu, Yuan Yao, Wendi Hou, Miaomiao Cui, Xuansong Xie, Changshui Zhang, and Xiansheng Hua. Boosting semantic human matting with coarse annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8563–8572, 2020.

[26] Qinglin Liu, Xiaoqian Lv, Quanling Meng, Zonglin Li, Xiangyuan Lan, Shuo Yang, Shengping Zhang, and Liqiang Nie. Revisiting context aggregation for image matting. *arXiv preprint arXiv:2304.01171*, 2023.

[27] Jingfeng Yao, Xinggang Wang, Lang Ye, and Wenyu Liu. Matte anything: Interactive natural image matting with segment anything model. *Image and Vision Computing*, 147:105067, 2024.

[28] Xiaodi Li, Zongxin Yang, Ruijie Quan, and Yi Yang. Drip: Unleashing diffusion priors for joint foreground and alpha prediction in image matting. *Advances in Neural Information Processing Systems*, 37:79868–79888, 2024.

[29] Zhixiang Wang, Baiang Li, Jian Wang, Yu-Lun Liu, Jinwei Gu, Yung-Yu Chuang, and Shin'Ichi Satoh. Matting by generation. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024.

[30] Kwanyong Park, Sanghyun Woo, Seoung Wug Oh, In So Kweon, and Joon-Young Lee. Mask-guided matting in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1992–2001, 2023.

[31] Yiqing Shi, Yiren Song, and Mike Zheng Shou. Edit2perceive: Image editing diffusion models are strong dense perceivers. *arXiv preprint arXiv:2511.18673*, 2025.

[32] Beomyoung Kim, Chanyong Shin, Joonhyun Jeong, Hyungsik Jung, Se-Yun Lee, Sewhan Chun, Dong-Hyun Hwang, and Joonsang Yu. Zim: Zero-shot image matting for anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23828–23838, 2025.

[33] Longfei Huang, Yu Liang, Hao Zhang, Jinwei Chen, Wei Dong, Lunde Chen, Wanyu Liu, Bo Li, and Peng-Tao Jiang. Sdmatte: Grafting diffusion models for interactive matting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15229–15239, 2025.

[34] Ruihao Xia, Yu Liang, Peng-Tao Jiang, Hao Zhang, Qianru Sun, Yang Tang, Bo Li, and Pan Zhou. Semat: Semantic enhanced natural image interactive matting. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[35] Xuecheng Li and Yuanjie Zheng. Inpaint-outpaint synergy: Mask refinement for trimap-free matting. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[36] He Guo, Zixuan Ye, Zhiguo Cao, and Hao Lu. In-context matting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3711–3720, 2024.

[37] Chenyi Zhang, Yihan Hu, Henghui Ding, Humphrey Shi, Yao Zhao, and Yunchao Wei. Learning trimaps via clicks for image matting. *IEEE Transactions on Multimedia*, 2025.

[38] Bo Xu, Han Huang, Cheng Lu, Ziwen Li, and Yandong Guo. Virtual multi-modality self-supervised foreground matting for human-object interaction. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 428–437. IEEE, 2021.

[39] Laurent D Cohen and Ron Kimmel. Global minimum for active contour models: A minimal path approach. *International journal of computer vision*, 24(1):57–78, 1997.

[40] Da Chen, Laurent D Cohen, Jean-Marie Mirebeau, and Xue-Cheng Tai. An elastica geodesic approach with convexity shape prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6900–6909, 2021.

[41] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. In *Proceedings of the 37th International Conference on Machine Learning*, pages 3789–3799, 2020.

[42] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021.

[43] Shaofei Wang, Bozidar Antic, Andreas Geiger, and Siyu Tang. Intrinsicavatar: Physically based inverse rendering of dynamic humans from monocular videos via explicit ray tracing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1877–1888, 2024.

[44] Fangzhou Mu, Carter Sifferman, Sacha Jungerman, Yiquan Li, Mark Han, Michael Gleicher, Mohit Gupta, and Yin Li. Towards 3d vision with low-cost single-photon cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5302–5311, 2024.

[45] Alejandro García-Castellanos, David R Wessels, Nicky J Berg, Remco Duits, Daniël M Pelt, and Erik J Bekkers. Equivariant eikonal neural networks: Grid-free, scalable travel-time prediction on homogeneous spaces. *arXiv preprint arXiv:2505.16035*, 2025.

[46] Jizhizi Li, Jing Zhang, and Dacheng Tao. Deep automatic natural image matting. *arXiv preprint arXiv:2107.07235*, 2021.

[47] Jizhizi Li, Jing Zhang, Stephen J Maybank, and Dacheng Tao. Bridging composite and real: towards end-to-end deep image matting. *International Journal of Computer Vision*, 130(2):246–266, 2022.

[48] Jizhizi Li, Sihan Ma, Jing Zhang, and Dacheng Tao. Privacy-preserving portrait matting. In *Proceedings of the 29th ACM international conference on multimedia*, pages 3501–3509, 2021.

[49] Jizhizi Li, Jing Zhang, and Dacheng Tao. Referring image matting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22448–22457, 2023.

[50] Christoph Rhemann, Carsten Rother, Jue Wang, Margrit Gelautz, Pushmeet Kohli, and Pamela Rott. A perceptually motivated online benchmark for image matting. In *2009 IEEE conference on computer vision and pattern recognition*, pages 1826–1833. IEEE, 2009.

[51] Samuel G Armato III, Geoffrey McLennan, Michael F McNitt-Gray, Charles R Meyer, David Yankelevitz, Denise R Aberle, Claudia I Henschke, Eric A Hoffman, Ella A Kazerooni, Heber MacMahon, et al. Lung image database consortium: developing a resource for the medical imaging research community. *Radiology*, 232(3):739–748, 2004.

[52] Lin Wang, Lie Ju, Donghao Zhang, Xin Wang, Wanji He, Yelin Huang, Zhiwen Yang, Xuan Yao, Xin Zhao, Xiufen Ye, et al. Medical matting: a new perspective on medical segmentation with uncertainty. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 573–583. Springer, 2021.

A Additional Implementation Details

Training setup. All experiments reported in this paper are conducted on the Huawei Cloud AI Compute Service. For reference, FreeMatte can also be trained locally on a single NVIDIA GeForce RTX 5090 D GPU with 32 GB of GPU memory and a batch size of 8, taking approximately two hours for one training run on AM-2K. We adopt ViTMatte [6] as the alpha prediction network. SAM-B [12] is used as a frozen segmentation foundation model to provide coarse localization priors; its parameters are not updated during training. Unless otherwise specified, all training settings not explicitly modified in our method, including the optimizer, learning-rate schedule, and number of training epochs, follow DDC-Matting [8].

Procedure. Algorithm 1 summarizes the training procedure of FreeMatte. For each augmented image and prior pair (I, S), ViTMatte predicts an alpha matte $\hat{\alpha}$. In parallel, the resistance function is constructed from the same pair and used to solve the Eikonal equation, producing an arrival-time field T. The field is then converted into per-pixel background sampling probabilities p_b, from which a sparse background mask M_b is sampled.

Three losses are used to train the matting network. The foreground loss $\mathcal{L}_{\text{fg}}$ encourages alpha values of one within the foreground region indicated by S. The background loss $\mathcal{L}_{\text{bg}}^{T}$ encourages alpha values of zero at pixels selected by M_b. The consistency loss $\mathcal{L}_{\text{ddc}}^{T}$ constrains alpha differences between visually similar neighboring pixels, with the arrival-time field modulating the target differences to suppress texture-induced variations in confident background regions.

Each image is revisited across training epochs, with a new background mask M_b sampled on every visit. Consequently, pixels with higher background sampling probabilities are more frequently selected as background anchors. The three losses are combined to update ViTMatte. Importantly, neither the arrival-time field nor the sampling probabilities serve as target alpha values; they are used only to construct and modulate the supervisory signals.

At inference, the frozen SAM-B generates a coarse localization prior from the input image and user prompts. ViTMatte then takes the image and prior as input and directly predicts the alpha matte, without requiring Eikonal field construction.

Algorithm 1 Training Procedure of FreeMatte

Input: Image–prior dataset $\mathcal{D}$; matting network M_θ with parameters θ.
for each epoch **do**
  for each mini-batch $\mathcal{B}$ from $\mathcal{D}$ **do**
    $\mathcal{L}_{\mathcal{B}} \leftarrow 0$
    for each image–prior pair (I, S) in $\mathcal{B}$ **do**
      $(I, S) \leftarrow \textsc{Augment}(I, S)$ {Data augmentation}
      $f \leftarrow \textsc{Resistance}(I, S)$ {Construct resistance from image cues}
      $T \leftarrow \textsc{FastMarching}(f, S)$ {Compute arrival-time field}
      $p_b \leftarrow \textsc{BackgroundProbability}(T)$ {Compute background sampling probabilities}
      $M_b(x) \sim \text{Bernoulli}(p_b(x)),\ \forall x$ {Sample background anchors}
      $\hat{\alpha} \leftarrow M_\theta(I, S)$
      $\mathcal{L}_{\text{fg}} \leftarrow \textsc{ForegroundLoss}(\hat{\alpha}, S)$ {Anchor foreground alpha to one}
      $\mathcal{L}_{\text{bg}}^{T} \leftarrow \textsc{BackgroundLoss}(\hat{\alpha}, M_b)$ {Anchor sampled background alpha to zero}
      $\mathcal{L}_{\text{ddc}}^{T} \leftarrow \textsc{GatedConsistencyLoss}(\hat{\alpha}, I, T)$ {Modulate DDC using T}
      $\mathcal{L}_{\mathcal{B}} \leftarrow \mathcal{L}_{\mathcal{B}} + \mathcal{L}_{\text{fg}} + \mathcal{L}_{\text{bg}}^{T} + \lambda_{\text{c}} \mathcal{L}_{\text{ddc}}^{T}$
    end for
    Backpropagate $\mathcal{L}_{\mathcal{B}}/|\mathcal{B}|$ and update θ.
  end for
end for
return M_θ

Hyperparameters. The consistency loss weight is set to $\lambda_{\text{c}} = 10$. For Eikonal-field construction, we use $\gamma = 10.0$, $\eta = 0.8$, and $r = 0.1$ in the resistance function. For sparse background anchoring, the arrival-time threshold is set to $\tau = 0.8$. For the Eikonal-gated DDC loss, the gate attenuation strength is set to $\kappa = 0.5$. For $\mathcal{L}_{\text{ddc}}^{T}$, we follow DDC-Matting to construct $\mathcal{N}(i)$. Specifically, candidate pixels within a $w \times w$ local window centered at pixel i are ranked by RGB distance, and

the top-K most similar pixels are selected. We set $w = 11$ and $K = 11$ in all experiments. These hyperparameters are fixed across all experiments unless otherwise stated.

Code. Our code is available at `https://github.com/deaodaoo/FreeMatte`. GT alpha mattes and trimaps are not used as prediction targets or loss supervision. In the released code, a GT-derived trimap is used solely to sample training crop centers from its unknown region; it is not provided to the prediction network or SAM-B, or used to construct the Eikonal field. Replacing this crop selector with randomly sampled SAM-B mask boundary pixels yields comparable MSE and Grad and lower SAD, MAD, and Conn on AM-2K (Table 5).

Table 5: Effect of the training crop-selection strategy on AM-2K. Lower values are better for all metrics.

Crop selector	MSE ↓	SAD ↓	MAD ↓	Grad ↓	Conn ↓
GT-derived trimap	0.008900	27.49	0.01600	17.43	11.47
SAM-B mask boundary	0.008950	25.90	0.01508	17.79	11.37

B Expectation and Stability of Sparse Supervision

In Sec. 3.4, the sparse background mask is sampled as

$$M_b(x) \sim \mathrm{Bernoulli}(p_b(x)), \tag{11}$$

where $p_b(x) \in [0, 1]$ is the background sampling probability derived from the Eikonal arrival-time field. This stochastic sampling admits a simple expectation-based interpretation.

Let $\ell_{\mathrm{bg}}(x) = |\hat{\alpha}(x)|$ denote the per-pixel background supervision term at location x. Ignoring the normalization factor used to stabilize the loss scale, the sampled background supervision can be written as:

$$\widetilde{\mathcal{L}}_{\mathrm{bg}}^{\mathrm{sample}} = \sum_{x \in \Omega} M_b(x)\, \ell_{\mathrm{bg}}(x). \tag{12}$$

Taking expectation with respect to the Bernoulli sampling process gives:

$$\begin{aligned} \mathbb{E}\Big[\widetilde{\mathcal{L}}_{\mathrm{bg}}^{\mathrm{sample}}\Big] &= \mathbb{E}\left[\sum_{x \in \Omega} M_b(x)\, \ell_{\mathrm{bg}}(x)\right] \\ &= \sum_{x \in \Omega} \mathbb{E}[M_b(x)]\, \ell_{\mathrm{bg}}(x). \end{aligned} \tag{13}$$

Since $M_b(x) \sim \mathrm{Bernoulli}(p_b(x))$, we have:

$$\mathbb{E}[M_b(x)] = p_b(x), \tag{14}$$

and therefore:

$$\mathbb{E}\Big[\widetilde{\mathcal{L}}_{\mathrm{bg}}^{\mathrm{sample}}\Big] = \sum_{x \in \Omega} p_b(x)\, \ell_{\mathrm{bg}}(x). \tag{15}$$

Eq. (15) shows that the numerator of the proposed stochastic sparse supervision is equivalent in expectation to a confidence-weighted background loss. Pixels with larger arrival-time scores are sampled more frequently and thus contribute more strongly to training, while ambiguous near-boundary pixels are naturally downweighted.

Sampling stability. We further analyze the stability of the stochastic background sampling process. Let

$$N_b = \sum_{x \in \Omega} M_b(x) \tag{16}$$

denote the number of sampled background anchors in one image. Since each $M_b(x)$ is sampled from a Bernoulli distribution with probability $p_b(x)$, we have:

$$\mathbb{E}[N_b] = \sum_{x \in \Omega} p_b(x), \qquad \mathrm{Var}(N_b) = \sum_{x \in \Omega} p_b(x)(1 - p_b(x)). \tag{17}$$

Because $p_b(x) \in [0,1]$, each Bernoulli variance satisfies $p_b(x)(1-p_b(x)) \le 1/4$, and therefore:

$$\mathrm{Var}(N_b) \le \frac{|\Omega|}{4}. \tag{18}$$

This shows that the fluctuation of the number of sampled anchors is bounded by the image size. Moreover, the sampling probabilities are explicitly controlled by the normalized arrival-time field and the threshold τ, so the sampling process does not introduce uncontrolled supervision density.

In practice, the background loss in Eq. (8) is normalized by $\sum_{x\in\Omega} M_b(x) + \varepsilon$. This normalization stabilizes the loss scale across different stochastic samples, while the Bernoulli process preserves the expectation of a confidence-weighted background supervision signal. Therefore, the proposed sparse supervision can be interpreted as a stochastic approximation to arrival-time-weighted background anchoring, while maintaining a stable loss scale during batch-based training.

C Additional Prior Analysis

Figure 7 compares the raw localization priors with the alpha mattes predicted by FreeMatte across different datasets and prior types. FreeMatte consistently reduces Grad in these comparisons, indicating improved boundary quality over the raw binary priors. Improvements in SAD and other region-level metrics are less uniform, as large foreground false positives or missing foreground regions can persist in the predictions. The final matting accuracy therefore remains dependent on the quality of the input prior.

Gradient error analysis. For point-prompted SAM-B priors, FreeMatte reduces Grad from 67.42 to 17.43 on AM-2K, from 92.61 to 35.80 on AIM-500, and from 76.24 to 20.30 on P3M-500-NP, corresponding to reductions of 74.1%, 61.3%, and 73.4%, respectively. Here, boundary awareness describes how image edges guide supervision construction; the continuous alpha matte is still learned from foreground and background anchors together with local consistency constraints.

Despite these improvements over the raw priors, a gap to stronger fully supervised methods remains. Foreground and background anchors constrain the alpha endpoints, while DDC constrains local alpha differences; these constraints do not uniquely determine continuous alpha transitions. Moreover, RGB differences can diverge from opacity differences under strong texture, similar foreground and background colors, adjacent objects, or internal transparency. Dense alpha supervision directly observes these transitions, whereas FreeMatte must infer them from indirect constraints.

To examine the residual errors, we analyze prior-quality subgroups and challenging conditions. The subgroup and threshold analyses below use a separately trained checkpoint with the same configuration; their aggregate scores may therefore differ slightly from the main results. As shown in Table 6, inaccurate priors are associated with substantially higher residual Grad across all three datasets.

Table 6: FreeMatte Grad error grouped by prior quality. Parentheses report sample counts. The increase is measured relative to the reliable-prior group.

Dataset	Reliable prior	Inaccurate prior	Increase
AM-2K	14.59 (142)	25.20 (58)	72.7%
AIM-500	25.64 (251)	47.91 (249)	86.9%
P3M-500-NP	15.84 (314)	27.56 (186)	74.0%

Table 7 further reports results for six manually tagged, non-exclusive failure conditions on AIM-500. FreeMatte improves Grad over SAM in every condition, but substantial residual errors remain, particularly for internal transparency and large holes. These subsets describe observed failure conditions, rather than exhaustive scene annotations, and an image may belong to multiple subsets. The results identify conditions associated with large residual errors. They do not provide a causal decomposition of the gap to supervised methods, for which per-image predictions are unavailable.

Connectivity error analysis. Figure 7 also shows that FreeMatte can obtain higher Conn errors than the raw binary priors. A binary localization mask and a continuous alpha matte represent

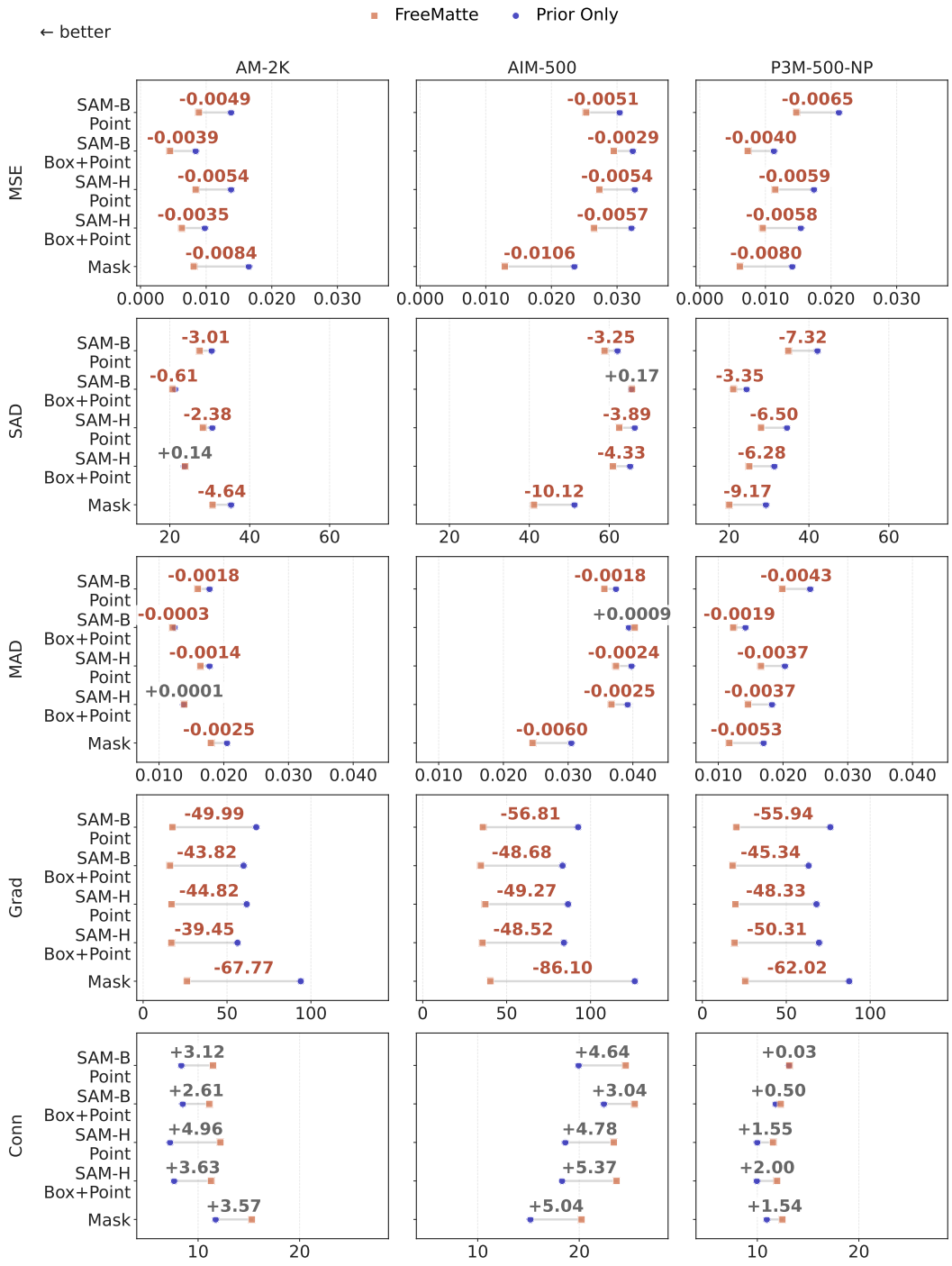

Figure 7: Full prior source analysis on AM-2K, AIM-500, and P3M-500-NP. Each subplot compares the raw localization prior (*Prior Only*) with the final alpha prediction produced by FreeMatte under the same prior source and prompting strategy. Columns correspond to datasets, and rows correspond to evaluation metrics. The horizontal axis reports error values, where lower is better. The value annotated between two endpoints indicates the error change from *Prior Only* to FreeMatte.

foreground structure differently. Binary masks can form compact, connected foreground regions, while continuous predictions include soft transitions, thin structures, and low-opacity regions that can change the measured connectivity.

Table 7: Errors under manually tagged challenging conditions on AIM-500. The subsets are non-exclusive, and n denotes the number of images in each subset.

Condition	n	Grad ↓		Conn ↓	
		SAM	FreeMatte	SAM	FreeMatte
Cluttered background	16	73.97	23.56	2.16	8.43
Adjacent/multiple objects	26	122.44	51.15	21.88	26.24
Similar foreground/background color	16	120.64	44.52	2.99	6.77
Internal transparency	17	201.37	97.62	126.48	130.85
Inaccurate mask prior	249	121.90	47.91	25.07	30.01
Large holes	7	406.86	249.81	70.99	74.27

To assess the sensitivity of Conn to output representation, we binarize the same FreeMatte predictions at thresholds $t = 0.5$ and $t = 0.7$. Table 8 shows that binarization substantially changes the Conn scores. At $t = 0.5$, it reduces the Conn gap relative to SAM by approximately 74% on AIM-500 and 87.8% on AM-2K. At $t = 0.7$, the binarized predictions obtain lower Conn than SAM on both datasets.

Table 8: Sensitivity of Conn to output binarization. Both thresholded outputs are obtained from the same continuous FreeMatte predictions. Lower values are better.

Output	AIM-500	AM-2K
SAM mask	19.94	8.35
FreeMatte alpha	23.42	11.22
FreeMatte mask ($t = 0.5$)	20.85	8.70
FreeMatte mask ($t = 0.7$)	19.68	7.41

These results indicate that Conn is substantially sensitive to output binarization, so its comparison between binary priors and continuous alpha mattes should be interpreted alongside the other matting metrics. The threshold analysis does not establish that all residual errors arise from representation differences. Incorrect foreground support and fragmented low-opacity structures can still introduce genuine connectivity errors.

D Data Scaling

To investigate the effect of additional training data, we augment the AM-2K training set with 50% and 100% of the P3M-10K training data [48]. The resulting training sets contain 6,510 and 11,221 images, respectively, compared with 1,800 images in AM-2K alone. All three models are evaluated on the AM-2K test set.

Table 9: Effect of adding P3M-10K training data to AM-2K. All results are evaluated on AM-2K. Lower values are better for all metrics.

Training data	Images	MSE ↓	SAD ↓	MAD ↓	Grad ↓	Conn ↓
AM-2K	1,800	0.0089	27.49	0.0160	17.43	11.47
AM-2K + 50% P3M	6,510	0.0091	25.78	0.0150	17.90	11.10
AM-2K + 100% P3M	11,221	0.0091	24.56	0.0143	18.14	10.15

As shown in Table 9, SAD, MAD, and Conn decrease monotonically as more P3M data is added. With the full P3M training set, SAD decreases from 27.49 to 24.56, while Conn decreases from 11.47 to 10.15. However, MSE increases slightly from 0.0089 to 0.0091, and Grad increases from 17.43 to 18.14. These results show that additional P3M training data benefits several metrics, although the improvements are not consistent across all metrics.

E Exploratory Cross-Domain Validation on Medical Images

To probe the cross-domain generality of FreeMatte, we conduct an additional experiment on the LIDC-IDRI [51] lung nodule dataset, which was re-annotated with matting-level alpha mattes by Medical Matting [52]. The evaluated images are 128×128 CT patches.

We use SAM3 [20] with the text prompt "pulmonary nodule" to generate localization priors, and compare FreeMatte with the raw SAM3 mask prior. We also include Closed-form matting as a traditional reference. Following Medical Matting [52], Closed-form uses a manual-mask-derived trimap, while FreeMatte does not use alpha mattes or trimaps during training. The alpha mattes are used only for evaluation.

Table 10: Cross-domain validation on LIDC-IDRI.

Method	Matting Ann.	MSE	SAD	MAD	Conn	Grad
Closed-form	manual trimap	0.4736	0.3040	–	0.3188	0.7584
SAM3 mask prior	none	0.0112	**0.2959**	**0.0181**	0.0494	0.7105
FreeMatte	none	**0.0076**	0.3317	0.0202	**0.0480**	**0.3523**

As shown in Table 10, FreeMatte improves the raw SAM3 prior on MSE, Conn, and Grad. The largest gain is on Grad, which is reduced from 0.7105 to 0.3523, suggesting that FreeMatte converts the hard binary prior into smoother boundary transitions.

FreeMatte is worse on SAD and MAD, which are region-accumulation metrics and are nearly linearly correlated at fixed resolution. This indicates mild alpha expansion around uncertain lesion boundaries. Overall, the results suggest that FreeMatte can transfer to low-resolution medical images and improve boundary-sensitive matting quality, while region-level accumulation remains a limitation when the input prior already provides strong object coverage.

F Limitations

Table 11: Counts of manually tagged failure conditions. Categories are non-exclusive, so an image may appear in multiple rows. These counts describe observed failures rather than exhaustive scene annotations. A zero indicates that no failure was tagged in that category during this audit.

Condition	AM-2K	AIM-500	P3M-500-NP
Cluttered background	5	16	1
Adjacent/multiple objects	3	26	10
Similar colors	2	16	0
Internal transparency	0	17	0
Inaccurate mask prior	58	249	186
Large holes	0	7	0

Our evaluation covers natural image matting benchmarks and exploratory medical-image validation. To examine the remaining limitations, we analyze six manually tagged failure conditions across three natural image datasets. Table 11 summarizes their counts, and Figure 8 presents representative examples. These conditions are non-exclusive and describe observed failures rather than exhaustive scene annotations. The quantitative analysis of residual Grad and Conn errors is provided in Appendix C.

Cluttered background. In cluttered scenes, background texture and edges can resemble foreground structures, making image-derived constraints less informative for foreground–background separation. FreeMatte may therefore retain spurious alpha responses in background regions.

Adjacent or multiple objects. Adjacent or overlapping objects can introduce ambiguity in the intended foreground support. When the localization prior includes unwanted objects or misses parts of the target object, boundary refinement alone may be insufficient to recover the intended foreground.

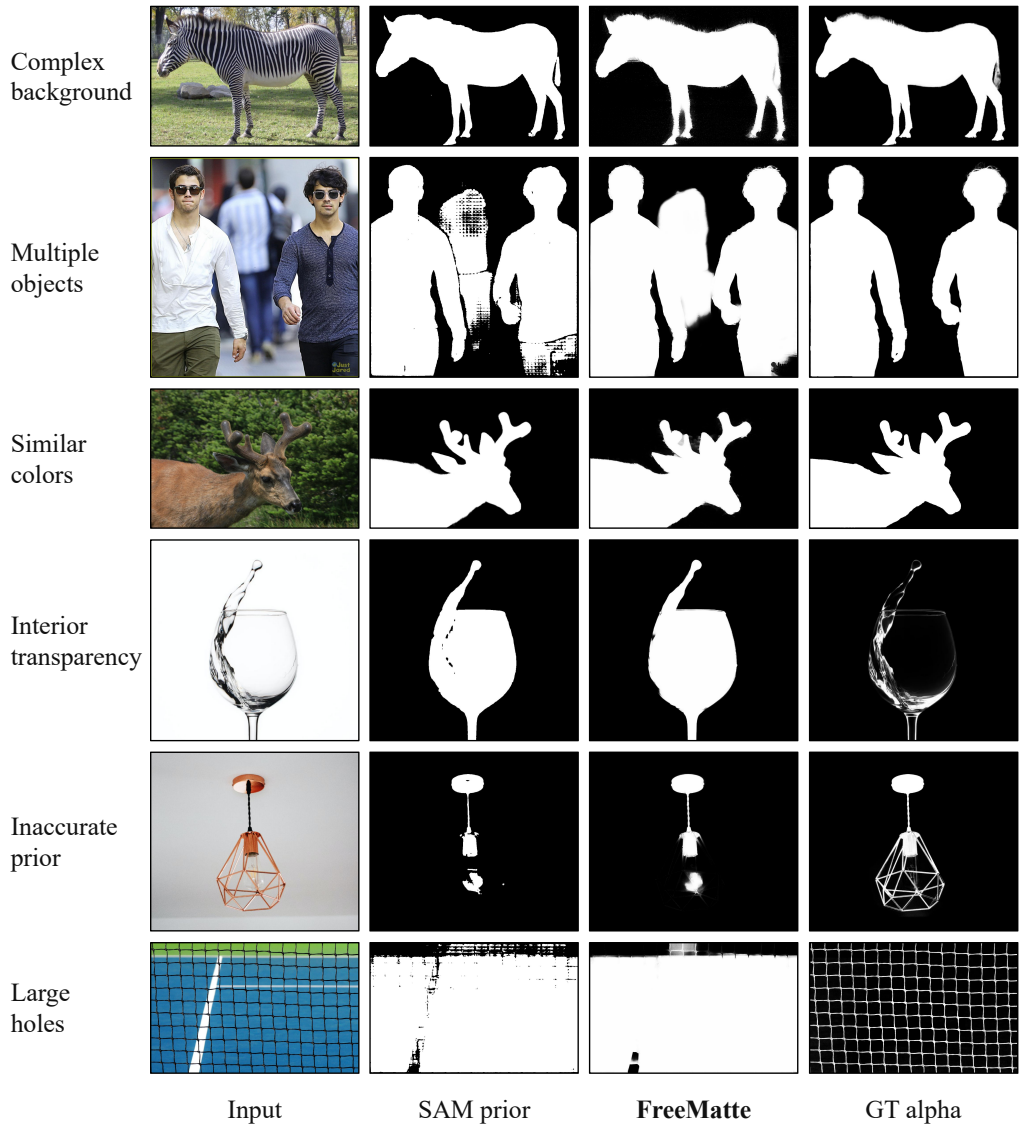

Figure 8: Representative failure cases of FreeMatte. Each row presents one example, with columns showing the input image, SAM-B prior, FreeMatte prediction, and ground-truth alpha matte, respectively. The examples illustrate observed limitations and do not indicate their frequency across datasets.

Similar foreground and background colors. RGB similarity does not necessarily imply similar opacity. When foreground and background colors are similar, appearance-based propagation and local consistency constraints may provide insufficient evidence for separating them and recovering accurate alpha transitions.

Internal transparency. A coarse binary prior does not encode spatially varying opacity within an object. Foreground and background anchors, together with local consistency constraints, may therefore be insufficient to recover complex internal transparency.

Inaccurate mask prior. Errors in the localization prior can propagate to the predicted alpha matte through the constructed training constraints. Foreground false positives may introduce incorrect foreground anchors, while omitted foreground structures may receive incorrect background constraints. In particular, thin structures missing from the prior can be suppressed if they receive high background

confidence and are sampled as background anchors. Inaccurate mask priors are the most frequently tagged failure condition across all three audited datasets.

Large holes. Objects with large holes require recovery of internal background regions as well as the outer object boundary. If these regions are incorrectly included in the foreground prior, the resulting constraints may hinder recovery of the correct foreground structure.

G Broader Impacts

FreeMatte may reduce the need for costly matting-specific annotations and improve the accessibility of image matting in annotation-scarce domains, such as medical imaging and other specialized applications. However, improved image matting can also be misused for deceptive image editing, unauthorized foreground extraction, or privacy-invasive manipulation of portrait images. Responsible use should respect dataset licenses, privacy expectations, and appropriate disclosure or provenance mechanisms when deployed in image-editing pipelines.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the paper's main contributions and scope, including the task-specific annotation-free matting setting, the Eikonal-field-based supervision framework, and evaluation on multiple benchmarks. The claims are supported by the method and experiments, and the paper makes clear that it avoids matting-specific annotations rather than all forms of annotations.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper includes a dedicated Limitations section discussing the dependence on coarse localization priors, limited modeling of complex internal transparency, and the scope of evaluation. It also notes that severe prior errors may not be recovered and that broader domains, transparent objects, and diverse segmentation priors require further study.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The paper does not claim broad theoretical guarantees; its theoretical analysis is limited to the expectation and stability interpretation of the sparse supervision. The assumptions, including Bernoulli sampling with probabilities derived from the arrival-time field, are stated, and the corresponding derivations are provided in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper discloses the key information needed to reproduce the main experimental results, including model choices, prior generation protocol, datasets, metrics, loss functions, hyperparameters, and implementation details. It also specifies that training settings not modified by the proposed method follow the referenced DDC-Matting protocol.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The submission provides code and instructions for reproducing the main experiments, together with details on data preparation, model configuration, hyperparameters, and evaluation protocols. The experiments use public datasets and pretrained models.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the main experimental settings, including the prediction network, segmentation prior model, prompt protocol, datasets, metrics, loss weights, and Eikonal-related hyperparameters. Additional implementation details such as GPU, batch size, training protocol, and neighborhood settings for the consistency loss are provided in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

Justification: The paper reports results on fixed benchmark test sets and includes ablations and hyperparameter sensitivity analyses, but does not report error bars, confidence intervals, or statistical significance tests over multiple runs. The conclusions are therefore based on deterministic benchmark comparisons rather than formal statistical significance analysis.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No] **Justification:** The paper reports the computational platform used for the experiments and provides an additional local hardware reference, including GPU memory, batch size, and approximate training time for one AM-2K training run. However, it does not provide execution time estimates for all individual experiments or the total compute required. Therefore, complete compute-resource reporting is not provided.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research uses standard public benchmarks and pretrained models for image matting, and does not involve human-subject experiments, crowdsourcing, or personally identifiable data collection. We have reviewed the NeurIPS Code of Ethics and are not aware of any aspect of the work that violates it.

   Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses positive impacts such as reducing costly matting-specific annotation requirements and enabling annotation-scarce applications, as well as potential negative impacts including deceptive image editing and privacy-invasive foreground extraction.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release a high-risk model or dataset such as an image generator, pretrained language model, or scraped dataset. Any released code or checkpoints are for image matting research on standard public benchmarks and will be distributed subject to the corresponding dataset and model licenses.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We use publicly available datasets and pretrained models in our experiments. All existing assets are cited in the paper, and we follow their corresponding licenses and terms of use. We use these assets only for research purposes and do not redistribute the original datasets or model weights.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The submission includes code and related documentation to facilitate reproducibility. The released assets include usage instructions, experimental settings, and necessary information about licenses and limitations.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing experiments or research with human subjects. It uses existing public benchmarks and does not collect new participant data or require participant instructions or compensation.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing experiments or research with human subjects. Therefore, there are no study participants, participant risks, or IRB approval requirements applicable to this work.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development does not involve LLMs as an important, original, or non-standard component.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.