OrienDiffusion: Taming Diffusion Towards Fine-grained Portrait Matting

Zhengzheng Tian[1] , **Wei Ma**[1*] , **Hongbin Zha**[2,3]
[1]College of Computer Science, Beijing University of Technology
[2]Anqing Normal University
[3]Peking University
tianz@emails.bjut.edu.cn, mawei@bjut.edu.cn, zha@cis.pku.edu.cn

Abstract

Recently, diffusion models have been applied to reformulate the regressive portrait matting task as a gradual denoising-driven generative process, alleviating the inherent limitations of discriminative models via iterative error correction. However, due to the lack of high-quality ground-truth mattes, existing generative methods for portrait matting still suffer from severe detail degradation, especially in structurally delicate hair strands and pixels with ambiguous colors and context. To address these challenges, we propose OrienDiffusion, a diffusion-based framework tailored for fine-grained portrait matting. Specifically, we construct a Hair Orientation Field to conditionally guide the denoising process, constraining the generation toward authentic structural consistency in hair regions. Furthermore, instead of relying on overly encoded deep features or ambiguous original color cues as existing methods do, we develop a Pixel Color-anchored Local Structure Embedding approach, which models alpha matte transitions as local structural variations, achieving accurate alpha estimation in pixels with ambiguous colors or context. Extensive experiments on multiple portrait matting datasets demonstrate that OrienDiffusion achieves state-of-the-art performance in these matting challenging regions. Code is available at https://github.com/xtz0001/OrienDiffusion.

1 Introduction

Portrait matting [Chuang *et al.*, 2001] [Sun *et al.*, 2004] [Chen *et al.*, 2013] [Levin *et al.*, 2007], a technique estimating pixel-wise opacity to extract human foregrounds from complex backgrounds, is fundamental to image editing, online conferencing, live streaming, etc. While significant progress has been made in this field [Wang *et al.*, 2024b], fine-grained portrait matting remains a core challenge in real-world scenarios, especially for thin-structured hair strands and small ambiguous-color regions in foreground-background transition zones, where contextual information is confusing. More critically, fine-grained alpha matte annotations are notoriously difficult to obtain [Liu *et al.*, 2025], resulting in unreliable supervision that impedes the training of alpha matting models. As shown in Figure 1, the hair strands in the ground-truth data suffer from structural discontinuity, and small background regions are mislabeled as foreground.

Figure 1: Fine-grained portrait matting results: input images, foregrounds extracted using ground-truth alpha mattes, and foregrounds extracted using OrienDiffusion-predicted alpha mattes. The proposed OrienDiffusion generates high-fidelity details for hair strands and pixels with ambiguous colors and context, even under supervision with noisy ground-truth data.

Most existing portrait matting approaches adopt discriminative models based on Convolutional Neural Networks (CNNs) or Transformers [Yao *et al.*, 2024] to directly regress alpha mattes from input images. By leveraging multi-scale features [Chen *et al.*, 2018] [Sun *et al.*, 2021] and global attention mechanisms [Li and Lu, 2020], these methods can effectively capture the overall characteristics of portraits for matting; however, they often struggle with thin hair strands and small regions with ambiguous colors or context.

Recently, generative diffusion models [Chen *et al.*, 2023] [Fu *et al.*, 2024] [Ke *et al.*, 2024] [Zhou *et al.*, 2024] have been introduced into the matting task [Li *et al.*, 2024] [Wang *et al.*, 2024a], providing a new modeling perspective for alpha estimation. By formulating matting as a gradual denoising-driven stochastic generation process, diffusion models can progressively recover high-frequency details through coarse-to-fine reconstruction. Existing works either

*Corresponding author.

fine-tune large-scale pre-trained generative models such as Stable Diffusion [Rombach *et al.*, 2022] or train lightweight diffusion denoising networks from scratch [Hu *et al.*, 2024]. Although generative matting methods exhibit certain advantages over traditional discriminative approaches in structural generation, the inherent randomness of the diffusion sampling process makes it difficult to precisely constrain the evolution trajectory of alpha values, particularly under imperfect ground-truth supervision. Therefore, when dealing with hair regions or small regions with ambiguity, the denoising process is prone to structural drift and detail degradation.

To address these issues, we propose a novel generative portrait matting framework termed OrienDiffusion, which is built on a lightweight diffusion model trained from scratch. Departing from sole reliance on ground-truth supervision, we emphasize structural priors during the generation process. Specifically, we construct a Hair Orientation Field (HOF) and develop a HOF injection strategy to incorporate the field as an explicit geometric prior for guiding the diffusion denoising process. This prior constrains alpha matte generation toward structurally consistent and visually natural hair patterns, even with inaccurate ground-truth annotations. Furthermore, for pixels with ambiguous color and contextual information, excessive feature abstraction undermines local structural discriminability, especially under imprecise supervision conditions. To mitigate this, we design a Pixel Color-anchored Local Structure Embedding (PCLSE) method that models spatial opacity transitions as local intensity patterns derived from original pixel colors within neighborhoods. This method furnishes stable local structural constraints for alpha matte generation in such ambiguous regions. As shown in Figure 1, even with noisy ground-truth data, our proposed model yields results with coherent hair structures and correctly classifies small background regions with ambiguous colors and context.

The contributions of this paper are as follows:

- We propose OrienDiffusion, a novel generative portrait matting framework based on a lightweight diffusion model. It integrates hair orientation continuity and local structural patterns as dual explicit structural priors to constrain the alpha matte generation trajectory, addressing the core limitations of diffusion models in fine-grained portrait matting.
- We introduce Hair Orientation Field guidance into the diffusion denoising process, which effectively ensures structural coherency in challenging hair strand regions by leveraging directional geometric priors.
- We design a Pixel Color-anchored Local Structure Embedding method that incorporates local structural cues to capture fine-grained pixel relationships, thereby improving alpha estimation accuracy in color- and context-ambiguous regions.
- Extensive experiments validate that OrienDiffusion achieves state-of-the-art (SOTA) performance on portrait matting tasks, exhibiting superiority particularly in handling challenging hair strands and color/context-ambiguous pixels.

2 Related Work

2.1 Deep Portrait Matting

Discriminative deep learning-based methods [Sengupta *et al.*, 2020] [Zhong and Zharkov, 2024] [Yu *et al.*, 2021a] [Sun *et al.*, 2022] have long dominated portrait matting, aiming to directly regress alpha mattes from RGB images in an end-to-end manner while maintaining semantic consistency and fine structural details.

Existing methods can be divided into auxiliary-based and auxiliary-free approaches. Auxiliary-based methods such as MGMatting [Yu *et al.*, 2021b], GCA [Li and Lu, 2020] and IndexNet [Lu *et al.*, 2019] leverage explicit priors to constrain ambiguous regions, demonstrating strong capability in recovering fine structures when accurate auxiliary are provided. However, their performance is highly sensitive to auxiliary quality [Kim *et al.*, 2025], limiting practical applicability.

To improve automation, auxiliary-free methods including [Shen *et al.*, 2016], MODNet [Ke *et al.*, 2022] and P3M [Li *et al.*, 2021] take only RGB images as input and compensate for missing priors using semantic cues and multi-scale context aggregation. These methods typically combine global semantic parsing with local detail refinement to balance structural integrity and boundary accuracy, but still struggle in color-ambiguous or structurally complex regions.

A fundamental limitation of discriminative approaches lies in formulating matting as one-shot regression, without explicit mechanisms for iterative refinement or uncertainty-aware correction. Consequently, predictions tend to be spatially smoothed to reduce regression variance, leading to blurred transitions, broken hair strands, and limited robustness to complex unseen scenes.

2.2 Generative Image Matting

Generative models reformulate portrait matting by explicitly modeling the uncertainty of alpha mattes, offering a fundamentally different perspective from discriminative regression. Early GAN-based approaches [Lutz *et al.*, 2018] [Xu *et al.*, 2022] introduced adversarial supervision to regularize alpha distributions, but their limited ability to synthesize fine-grained structures and reliance on carefully balanced objectives hinder robustness in complex scenes.

Recently, diffusion models [Yu *et al.*, 2025] have emerged as a powerful generative paradigm for portrait matting, leveraging iterative denoising to capture high-frequency details and complex alpha distributions. By treating the alpha matte as a samplable distribution rather than a direct regression target, diffusion-based methods demonstrate improved robustness in challenging backgrounds and boundary regions. However, existing approaches [Huang *et al.*, 2025] [Li *et al.*, 2024] either rely on large pre-trained diffusion models [Rombach *et al.*, 2022] with prohibitive computational cost or adopt lightweight diffusion frameworks [Xu *et al.*, 2025] [Hu *et al.*, 2024] that treat denoising as post-hoc refinement with fixed trajectories. As a result, the generation process lacks explicit structural controllability [Liu *et al.*, 2024], particularly in hair regions where directional consistency and local continuity are critical [Zhang *et al.*, 2023] [Ye *et al.*, 2023], leading to detail degradation during diffusion.

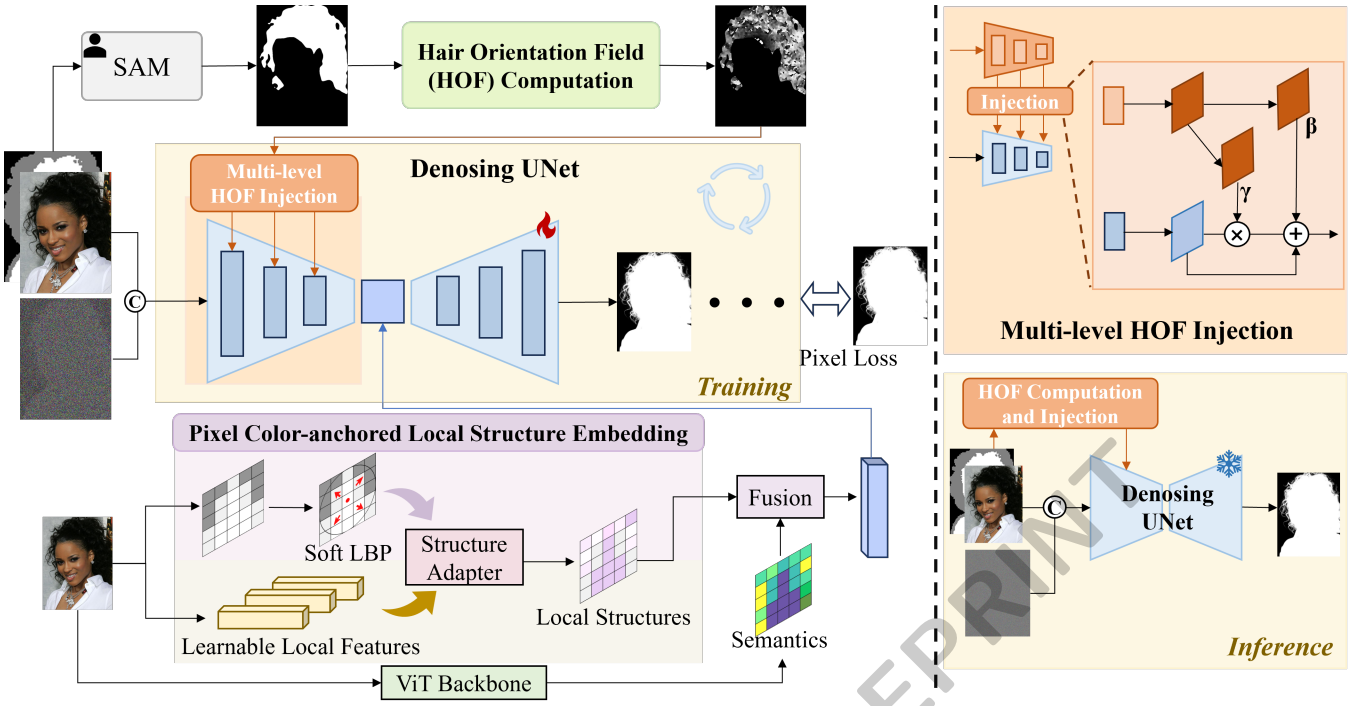

Figure 2: Overall pipeline of the proposed OrienDiffusion. We perform Hair Orientation Field Computation and Multi-level HOF Injection (Section 3.2) to guide the UNet denosing process. In addition, we design a Pixel Color-anchored Local Structure Embedding module (Section 3.3) to capture pixel-wise fine-grained patterns as denoising prior clues.

3 Method

3.1 Overview

We formulate portrait matting as a diffusion-based alpha generation problem, where the target alpha matte is progressively refined through a conditional denoising process. Specifically, given the ground-truth alpha matte x_0, the forward diffusion process gradually corrupts it by adding Gaussian noise [Ho *et al.*, 2020], producing a sequence of latent variables:

$$q(x_t \mid x_0) = \mathcal{N}\left(x_t; \sqrt{\bar{\alpha}_t}\, x_0,\ (1-\bar{\alpha}_t)\mathbf{I}\right) \tag{1}$$

where $t \in \{1, \ldots, T\}$ denotes the diffusion timestep, $\bar{\alpha}_t = \prod_{i=1}^{t}(1-\beta_i)$, β_i is a predefined noise schedule, and $\mathbf{I}$ denotes the identity matrix. During the reverse process, the denoising process is conditioned on the input image I, Trimap Tri, and additional structural priors C derived from I. Accordingly, we learn a conditional reverse denoising process:

$$p_\phi(x_{t-1} \mid x_t, I, Tri, C) \tag{2}$$

which progressively recovers the clean alpha matte from noisy samples.

As illustrated in Figure 2, firstly, a Hair Orientation Field is computed from I and injected into the diffusion process to provide geometry guidance for hair regions (Section 3.2). This orientation prior enforces directional continuity during denoising, guiding the model toward structurally coherent hair alpha generation. Secondly, for regions with severe color or context ambiguity, we design a Pixel Color-anchored Local Structure Embedding module (Section 3.3), which captures pixel-wise constraints based on local neighborhood comparisons. These priors are incorporated into the denoising network at each diffusion step, explicitly constraining the denoising trajectory and leading to more accurate and structurally consistent alpha matte estimation.

3.2 Hair Orientation Field Computation and Injection

Hair-Orientation Field Computation. Hair regions in portrait matting exhibit strong strand-level directional continuity. Explicitly modeling such directional cues provides valuable structural guidance for alpha estimation, helping preserve spatial coherence and maintain consistent hair patterns during the matting process.

We therefore propose a HOF-guided denoising strategy, where a continuous orientation field is estimated following [Tan *et al.*, 2020] [Pan *et al.*, 2025] and injected as structural prior to constrain the diffusion trajectory. The HOF enforces spatially consistent feature evolution along hair growth directions, enabling more coherent strand structures and improved matting of fine hair details.

Specifically, the construction of the HOF first relies on accurate hair region localization. In practical applications, portrait images are often not standard half-body photos, but are frequently accompanied by complex poses, occlusions, and accessories such as hats and hair ornaments, making it difficult to obtain a reliable hair mask using pre-trained portrait or hair segmentation models. To improve hair region extraction capabilities in complex scenes, we introduce the segmenta-

tion extraction capabilities of the SAM3 [Carion *et al.*, 2025]. Potential hair regions are coarsely located using text prompts (such as "hair of the people") to obtain a hair mask for selecting hair regions in the orientation field. For portraits that do not contain hair regions (e.g., wearing a hat or with no visible hair), we set their hair mask to a zero matrix to avoid introducing incorrect structural priors and ensure that the model maintains higher sensitivity to real hair samples during training.

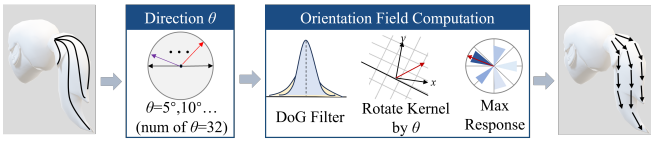

Figure 3: Illustration of the Hair Orientation Field computation.

The angle values of the HOF are determined based on the response of each pixel to different orientation filters, as illustrated in Figure 3. Specifically, we first convert image I to a grayscale image I_g. Subsequently, we use a set of differential Gaussian directional filters $\{K_\theta\}_{\theta=1}^{N}$ with $N = 32$, where the corresponding orientations are uniformly sampled from $[0, \pi)$, to compute the directional responses of the image:

$$R_\theta(p) = (K_\theta \otimes I_g)_p \tag{3}$$

where $\otimes$ denotes the convolution operator and $R_\theta(p)$ represents the response at pixel p under direction θ. To suppress non-structural noise, only positive responses are retained. The dominant orientation at pixel p is determined by the filter that yields the maximum response:

$$\theta_p = \arg\max_\theta R_\theta(p) \tag{4}$$

Meanwhile, the corresponding maximum response is used as the orientation confidence:

$$w_p = \max_\theta R_\theta(p) \tag{5}$$

Since the above procedure produces discrete orientation labels and hair orientation exhibits periodic symmetry (i.e., θ and $\theta + \pi$ are equivalent), we map the discrete orientations to continuous two-dimensional direction vectors:

$$\mathbf{O}_p = \left[\cos(2\theta_p),\ \sin(2\theta_p)\right] \tag{6}$$

The direction vector field is then Gaussian smoothed to suppress local noise and enhance overall direction consistency. Only the direction vectors within the hair mask are retained, and encoded with the confidence w_p into the feature vector.

Multi-level HOF Injection. We construct a HOF O to encode the continuous growth directions of hair strands. To effectively leverage this structural prior in diffusion-based matting, we introduce a multi-level HOF injection mechanism inspired by [Huang and Belongie, 2017] [Park *et al.*, 2019] [Zhang *et al.*, 2025] that aligns the orientation guidance with the hierarchical feature representations of the U-Net denoising network.

Specifically, given the HOF O, we employ a conditional encoder to extract a set of multi-scale orientation features F_o^l (l = 1, 2, 3) through progressive downsampling. These features are spatially aligned with the corresponding layers of the U-Net backbone, enabling orientation-aware constraints to be imposed throughout the diffusion denoising process. Low-level features capture fine-grained strand directions to stabilize local structures, while high-level features encode global hairstyle layouts to enforce long-range directional consistency.

To inject orientation cues into the diffusion network, we adopt a feature modulation scheme. For each U-Net feature F_u^l and its corresponding orientation feature F_o^l, a mapping network conditioned on F_o^l predicts a scale γ^l and a bias β^l, and the modulated feature is computed as

$$\hat{F}_u^l = (1 + \gamma^l) \odot F_u^l + \beta^l \tag{7}$$

Here, $\odot$ denotes element-wise multiplication, where γ^l enhances structure-aware responses in hair regions and β^l provides directional guidance without disrupting semantic representations. The residual form $1 + \gamma^l$ ensures stable feature propagation during training.

By injecting orientation priors across multiple scales in a controlled manner, the diffusion model preserves its denoising capability while being explicitly guided toward structurally coherent hair alpha generation. This multi-level HOF injection enables more continuous and visually consistent hair structures in complex portrait scenes.

3.3 Pixel Color-anchored Local Structure Embedding

In matting regions where foreground and background colors are highly similar, alpha estimation based on color cues becomes unreliable. Although existing methods enhance discrimination using high-level Transformer features [Hu *et al.*, 2024], alpha regression in portrait matting still fundamentally depends on relative intensity variations and local mixing relationships. Such fine-grained structural cues are often weakened by patch-based modeling and progressive semantic abstraction in Vision Transformers, and this issue is further aggravated by imperfect supervision, leading to blurred transition boundaries.

To mitigate this problem, we propose a Pixel Color-anchored Local Structure Embedding module that constrains the solution space using pixel colors as anchors while explicitly reinforcing local structural consistency. When color anchors become ambiguous, reliable alpha estimation must rely more on structural relationships within local neighborhoods. From the linear compositing model $I = \alpha F + (1 - \alpha)B$, the image gradient can be written as

$$\nabla I = \alpha \nabla F + (1 - \alpha)\nabla B + (F - B)\nabla \alpha \tag{8}$$

When F and B are indistinguishable in color space, the contrast term $(F - B)\nabla\alpha$ diminishes, causing the image gradient ∇I to be dominated by foreground and background textures, making color differences insufficient for inferring alpha variations. Nevertheless, local texture and structural patterns still implicitly encode the spatial transition of pixel mixing,

motivating us to constrain alpha prediction from a structural perspective. Specifically, we introduce Local Binary Pattern (LBP) structural cues, which capture relative intensity variations within local neighborhoods and are less sensitive to absolute color similarity. For a sampling direction ϕ on a circle of radius r, the difference response can be approximated as

$$D_{(r,\phi)}I(x) \approx I(x + r\mathbf{u}_\phi) - I(x), \mathbf{u}_\phi = (\cos\phi, \sin\phi) \quad (9)$$

where $\mathbf{u}_\phi$ denotes the unit vector along direction ϕ. This response reflects local intensity transitions and remains sensitive to alpha changes across unknown regions even under low-contrast conditions.

Since standard LBP is non-differentiable and scale-fixed [Pietikäinen, 2010] [Ahonen *et al.*, 2004], we design a continuous and learnable variant suitable for end-to-end training. For a gray scale image G, it is formulated as

$$\mathrm{LBP}_r(G) = \sum_{k=0}^{K-1} \sigma\left(\frac{G_k - G_c}{\tau}\right) 2^k \quad (10)$$

where G_c and G_k denote the center and neighboring pixels at scale r, $\sigma(\cdot)$ is the Sigmoid function, and τ is a learnable temperature parameter controlling the smoothness of binarization. Multi-scale LBP responses are averaged to obtain the fused structural representation.

To further enhance structural representation, we introduce a shallow convolutional branch to learn complementary local patterns and concatenate them with the LBP responses to construct the final structural features. These structural features are then adaptively fused with semantic features via attention weighting and injected into the diffusion bottleneck through lightweight adapters.

4 Experiments

4.1 Implementation Details

Datasets and Evaluation Metric. The P3M-10K dataset [Li *et al.*, 2021] is one of the most widely used portrait matting datasets, containing 9,421 high-resolution portrait images and alpha mattes. The HPM-17K dataset, introduced by MTGINet [Yang *et al.*, 2025], contains approximately 17,000 half-body images and mattes. Both datasets cover a wide range of human poses and diverse real-world scenes. We conduct training and evaluation on both datasets to assess the performance of our method.

We evaluate our model using commonly adopted matting metrics [Rhemann *et al.*, 2009] including Sum of Absolute Differences (SAD), Mean Squared Error (MSE), Gradient loss (Grad), and Connectivity loss (Conn). In regions where the ground-truth hair strands are inaccurate, we focus on generating visually coherent and naturally synthesized alpha mattes.

Training Details. Our conditional denoising network UNet consists of downsampling and upsampling convolutional blocks. Our model was trained for 500,000 steps on a single NVIDIA 4090 GPU with a batch size of 2 within 48 hours, and has 819 GFLOPs. For the loss function, we employ pixel-level loss composed of MSE, L1, gradient, and Laplacian losses followed [Hu *et al.*, 2024]. During inference, our model requires approximately 1 second to process a single image.

4.2 Ablation Study

Effect of Modules. Table 1 shows that the basic generative framework alone performs poorly across metrics. Adding HOF consistently improves performance, indicating that explicit structural constraints enhance alpha continuity. However, orientation cues alone are insufficient for distinguishing regions with similar colors. Incorporating PCLSE further boosts performance, and even alone yields significant gains, demonstrating its effectiveness in improving alpha estimation in such ambiguous regions.

HOF	PCLSE	SAD↓	MSE↓	Grad↓	Conn↓
		3.62	0.0028	4.12	3.07
√		3.32	0.0022	3.81	2.78
	√	**2.91**	<u>0.0018</u>	<u>2.99</u>	**2.39**
√	√	<u>2.96</u>	**0.0017**	**2.77**	<u>2.40</u>

Table 1: Ablation study on the key modules on the HPM-17K dataset.

Figure 4 shows the results of the visualization ablation experiment. It can be seen that when the HOF injection module is removed (third column), the structural continuity of the hair region decreases significantly and local blurring occurs. After removing PCLSE (second column), blurring and misjudgment are more likely to occur at the foreground-background boundary. In contrast, the complete model OrienDiffusion maintains a more stable and natural alpha transition in complex boundaries, verifying the effectiveness of each module in detail restoration and structural consistency.

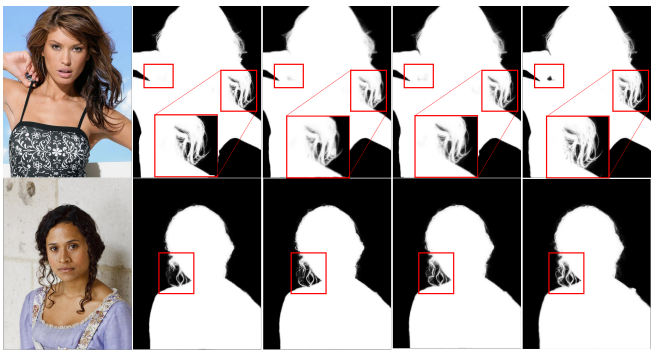

Figure 4: Visual ablation study on the key modules in OrienDiffusion.

Effect of Hyperparameter β. As shown in Table 2, we analyze the effect of the bias term β used in the HOF injection. The results indicate that the model achieves consistently better performance across all metrics when β is used jointly with γ, compared to using either term alone. This suggests that β, serving as an orientation-guided compensation term, can effectively introduce structural priors when added to the weighted features especially around boundary regions.

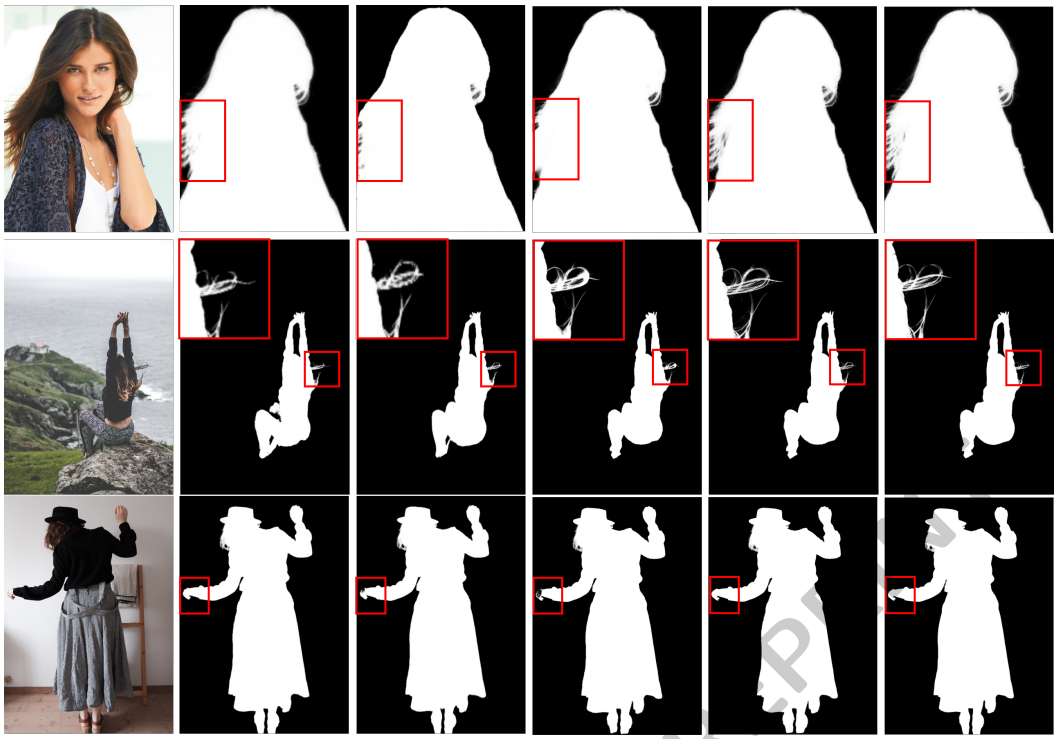

Figure 5: Qualitative comparison on the portrait matting dataset. As shown, our method better preserves fine-grained hair structures and obtains more accurate alpha values in color- and context-ambiguous regions.

Variants	SAD ↓	MSE ↓	Grad ↓	Conn ↓
w/o β	3.17	0.0021	3.40	2.62
w/ β	**2.96**	**0.0017**	**2.77**	**2.40**

Table 2: Effects of variant β in HOF injection on the HPM-17K dataset.

Effect of Injection Mechanisms. We compare our injection method with two strategies: ControlNet-style zero convolution and feature concatenation. As given in Table 3, our injection method performs well. Specifically, we perform fine-grained modulation on hair-region features while preserving feature coherence. In contrast, ControlNet-style conditioning and feature concatenation lack fine-grained modulation and may disrupt structural consistency across the modulation process.

Method	SAD ↓	MSE ↓
Base model	3.62	0.0028
+HOF w/ ControlNet	3.36	0.0173
+HOF w/ Concatenation	3.84	0.0226
+HOF w/ Our HOF injection	**3.32**	**0.0022**

Table 3: Comparison of HOF injection methods on the HPM-17K dataset.

Methods	SAD↓	MSE↓	Grad↓	Conn↓
MODNet [Ke *et al.*, 2022]	4.73	0.0037	4.24	3.45
P3MNet [Li *et al.*, 2021]	3.13	0.0024	<u>2.63</u>	2.41
DiffMatte [Hu *et al.*, 2024]	3.62	0.0028	4.12	3.07
MTGINet [Yang *et al.*, 2025]	<u>3.11</u>	<u>0.0021</u>	**2.28**	**2.18**
OrienDiffusion (Ours)	**2.96**	**0.0017**	2.77	<u>2.40</u>

Table 4: Quantitative comparison results on the HPM-17K dataset.

4.3 Comparative Experiments

Comparison on HPM-17K. We conduct comprehensive quantitative and qualitative comparisons with representative portrait matting methods, including MODNet [Ke *et al.*, 2022], P3MNet [Li *et al.*, 2021], MTGINet [Yang *et al.*, 2025], and DiffMatte [Hu *et al.*, 2024], covering both discriminative approaches and generative diffusion-based models. The quantitative results are summarized in Table 4. Our method achieves the best or second-best performance across the SAD, MSE, and Conn metrics, consistently outperforming existing approaches. These results demonstrate that the proposed OrienDiffusion is more effective in recovering continuous and natural alpha distributions, particularly in regions with fine-grained hair structures and severe foreground-background ambiguity.

It is worth noting that although MTGINet [Yang *et al.*, 2025] reports relatively low errors on certain metrics such as Grad, its visual results still fail to achieve hair-level preci-

Category	Methods	P3M-500-NP			P3M-500-P			Parameters (M)↓
		SAD↓	MSE↓	Conn↓	SAD↓	MSE↓	Conn↓	
Discriminative	P3MNet [Li *et al.*, 2021]	11.23	0.0035	12.51	8.73	0.0026	9.68	39.48
	MODNet [Ke *et al.*, 2022]	16.7	0.0051	13.81	13.11	0.0038	10.88	6.49
	ViT-Matte [Yao *et al.*, 2024]	7.26	0.0012	–	–	–	–	25.8
	MTGINet [Yang *et al.*, 2025]	12.71	0.0040	6.85	9.04	0.0026	9.73	29.72
Generative	AlphaLDM [Wang *et al.*, 2024a]	–	–	–	7.1	0.0016	6.8	–
	DiffMatte [Hu *et al.*, 2024]	6.66	0.0012	4.85	6.23	**0.0012**	4.53	29.0
	SDMatte [Huang *et al.*, 2025]	**5.1**	**0.0007**	**4.12**	–	–	–	957
	LiteSDMatte [Huang *et al.*, 2025]	6.66	0.0011	5.22	–	–	–	593
	OrienDiffusion (Ours)	5.93	0.0010	4.49	**5.69**	**0.0012**	**4.08**	31.8

Table 5: Quantitative comparison results on the P3M-10K dataset.

sion. As illustrated in the first row of Figure 5, structural discontinuities and hair breakage remain evident in fine-grained regions, indicating that its results struggle to preserve subtle hair structures. In contrast, despite potential discrepancies in quantitative evaluation, OrienDiffusion consistently produces more coherent and visually plausible alpha mattes, especially in delicate hair boundaries.

Comparison on P3M-10K. The P3M-10K dataset presents a more challenging evaluation setting, containing a larger proportion of full-body portraits captured in complex real-world scenes. P3M-500-NP, which consists of normal portraits without occlusions, and P3M-500-P, where facial regions are partially occluded by masks. Our method outperforms the discriminative matting models in terms of quantitative metrics, indicating a higher level of foreground-background separation quality. We further compare the model complexity of different approaches. As reported in Table 5, SDMatte [Huang *et al.*, 2025] achieves state-of-the-art quantitative performance but requires approximately 30× more parameters than our method. Its distilled variant, LiteSDMatte, still contains nearly 20× parameters while yielding inferior performance compared to OrienDiffusion.

Due to the generative nature of diffusion models, OrienDiffusion produces richer alpha details in high-frequency regions (e.g. hair). However, imperfect ground-truth annotations often fail to capture such fine structures, biasing quantitative evaluation; thus, metric gaps do not fully reflect detail recovery. As shown in the second and third rows of Figure 5, our method yields more natural and continuous alpha mattes on complex backgrounds, reduces hair breakage, and remains robust in challenging cases with similar foreground-background colors.

In addition, OrienDiffusion can be naturally extended to matting animal fur. We extract the unknown region of the animal trimap to extract the corresponding orientation field to guide inference. As shown in Figure 6, owing to the shared filamentary structures between human hair and animal fur, the model preserves directional consistency and produces smooth, coherent transitions in fine fur regions. Moreover, it maintains stable boundaries under complex backgrounds and varying illumination, demonstrating strong generalization capability beyond human hair.

Figure 6: Generalization results on animal matting: input images, alpha mattes predicted by OrienDiffusion, and compositions with new backgrounds using the predicted alpha mattes.

5 Conclusion

In summary, we proposed OrienDiffusion, a generative portrait matting framework that incorporates explicit structural priors to address two long-standing challenges in portrait matting: structural discontinuity in fine hair regions and severe color ambiguity between foreground and background. By formulating matting as a conditional diffusion process, our method benefits from iterative error correction, which is particularly effective for recovering delicate and spatially coherent alpha structures. Specifically, we introduced HOF as geometric guidance to enforce spatial continuity of hair strands during denoising, and designed a PCLSE module with LBP-based structural constraints to enhance robustness in regions with similar foreground and background colors. Extensive experiments demonstrate that OrienDiffusion shows promising generalization to furry objects such as animals.

Nevertheless, like most existing matting methods, our method still relies on imperfect ground-truth alpha mattes for supervision, which may bias training toward numerical fidelity over perceptual quality, especially in fine hair regions. Moreover, the high computational cost (819 GFLOPs) limits the inference speed and practical deployment. Future work will explore human-preference-aware quality assessment or perceptual feedback as auxiliary supervision signals, and lightweight designs to improve both quality and efficiency.

Acknowledgments

This work was supported by the Beijing Natural Science Foundation (No. 4252029) and the National Natural Science Foundation of China (No. 62176010).

References

[Ahonen *et al.*, 2004] Timo Ahonen, Abdenour Hadid, and Matti Pietikäinen. Face recognition with local binary patterns. In *European Conference on Computer Vision*, pages 469–481, 2004.

[Carion *et al.*, 2025] 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.

[Chen *et al.*, 2013] Qifeng Chen, Dingzeyu Li, and Chi-Keung Tang. Knn matting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(9):2175–2188, 2013.

[Chen *et al.*, 2018] Quan Chen, Tiezheng Ge, Yanyu Xu, Zhiqiang Zhang, Xinxin Yang, and Kun Gai. Semantic human matting. In *ACM International Conference on Multimedia*, pages 618–626, 2018.

[Chen *et al.*, 2023] Shoufa Chen, Peize Sun, Yibing Song, and Ping Luo. Diffusiondet: Diffusion model for object detection. In *IEEE International Conference on Computer Vision*, pages 19830–19843, 2023.

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

[Fu *et al.*, 2024] Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and Xiaoxiao Long. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In *European Conference on Computer Vision*, pages 241–258, 2024.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[Hu *et al.*, 2024] Yihan Hu, Yiheng Lin, Wei Wang, Yao Zhao, Yunchao Wei, and Humphrey Shi. Diffusion for natural image matting. In *European Conference on Computer Vision*, pages 181–199, 2024.

[Huang and Belongie, 2017] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *IEEE International Conference on Computer Vision*, pages 1501–1510, 2017.

[Huang *et al.*, 2025] 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 *IEEE International Conference on Computer Vision*, pages 15229–15239, 2025.

[Ke *et al.*, 2022] Zhanghan Ke, Jiayu Sun, Kaican Li, Qiong Yan, and Rynson WH Lau. Modnet: Real-time trimap-free portrait matting via objective decomposition. In *AAAI Conference on Artificial Intelligence*, pages 1140–1147, 2022.

[Ke *et al.*, 2024] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 9492–9502, 2024.

[Kim *et al.*, 2025] 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 *IEEE International Conference on Computer Vision*, pages 23828–23838, 2025.

[Levin *et al.*, 2007] 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.

[Li and Lu, 2020] Yaoyi Li and Hongtao Lu. Natural image matting via guided contextual attention. In *AAAI Conference on Artificial Intelligence*, pages 11450–11457, 2020.

[Li *et al.*, 2021] Jizhizi Li, Sihan Ma, Jing Zhang, and Dacheng Tao. Privacy-preserving portrait matting. In *ACM International Conference on Multimedia*, pages 3501–3509, 2021.

[Li *et al.*, 2024] 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*, pages 79868–79888, 2024.

[Liu *et al.*, 2024] Haipeng Liu, Yang Wang, Biao Qian, Meng Wang, and Yong Rui. Structure matters: Tackling the semantic discrepancy in diffusion models for image inpainting. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8038–8047, 2024.

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

[Lu *et al.*, 2019] Hao Lu, Yutong Dai, Chunhua Shen, and Songcen Xu. Indices matter: Learning to index for deep image matting. In *IEEE International Conference on Computer Vision*, pages 3266–3275, 2019.

[Lutz *et al.*, 2018] Sebastian Lutz, Konstantinos Amplianitis, and Aljosa Smolic. Alphagan: Generative adversarial networks for natural image matting. *arXiv preprint arXiv:1807.10088*, 2018.

[Pan *et al.*, 2025] Yimin Pan, Matthias Niessner, and Tobias Kirschstein. Hairgs: Hair strand reconstruction based on 3d gaussian splatting. *arXiv preprint arXiv:2509.07774*, 2025.

[Park *et al.*, 2019] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with

spatially-adaptive normalization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2337–2346, 2019.

[Pietikäinen, 2010] Matti Pietikäinen. Local binary patterns. *Scholarpedia*, 5(3):9775, 2010.

[Rhemann *et al.*, 2009] Christoph Rhemann, Carsten Rother, Jue Wang, Margrit Gelautz, Pushmeet Kohli, and Pamela Rott. A perceptually motivated online benchmark for image matting. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1826–1833, 2009.

[Rombach *et al.*, 2022] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[Sengupta *et al.*, 2020] Soumyadip Sengupta, Vivek Jayaram, Brian Curless, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Background matting: The world is your green screen. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2291–2300, 2020.

[Shen *et al.*, 2016] Xiaoyong Shen, Xin Tao, Hongyun Gao, Chao Zhou, and Jiaya Jia. Deep automatic portrait matting. In *European Conference on Computer Vision*, pages 92–107, 2016.

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

[Sun *et al.*, 2021] Yanan Sun, Chi-Keung Tang, and Yu-Wing Tai. Semantic image matting. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 11120–11129, 2021.

[Sun *et al.*, 2022] Yanan Sun, Chi-Keung Tang, and Yu-Wing Tai. Human instance matting via mutual guidance and multi-instance refinement. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2647–2656, 2022.

[Tan *et al.*, 2020] Zhentao Tan, Menglei Chai, Dongdong Chen, Jing Liao, Qi Chu, Lu Yuan, Sergey Tulyakov, and Nenghai Yu. Michigan: multi-input-conditioned hair image generation for portrait editing. *ACM Transactions on Graphics*, 39(4):95–1, 2020.

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

[Wang *et al.*, 2024b] Zitao Wang, Qiguang Miao, Yue Xi, and Peipei Zhao. Eformer: Enhanced transformer towards semantic-contour features of foreground for portraits matting. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3880–3889, 2024.

[Xu *et al.*, 2022] Yangyang Xu, Zeyang Zhou, and Shengfeng He. Self-supervised matting-specific portrait enhancement and generation. *IEEE Transactions on Image Processing*, 31:5332–5342, 2022.

[Xu *et al.*, 2025] Yangyang Xu, Shengfeng He, Wenqi Shao, Yong Du, Kwan-Yee K Wong, Yu Qiao, Jun Yu, and Ping Luo. Diffusionmat: Alpha matting as deterministic sequential refinement learning. In *ACM International Conference on Multimedia*, pages 9454–9462, 2025.

[Yang *et al.*, 2025] Wenbing Yang, Wei Ma, Qing Mi, and Hongbin Zha. Multi-task gradual inference with a single encoder–decoder network for automatic portrait matting. *Computational Visual Media*, 11(6):1385–1398, 2025.

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

[Ye *et al.*, 2023] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

[Yu *et al.*, 2021a] Haichao Yu, Ning Xu, Zilong Huang, Yuqian Zhou, and Humphrey Shi. High-resolution deep image matting. In *AAAI Conference on Artificial Intelligence*, volume 35, pages 3217–3224, 2021.

[Yu *et al.*, 2021b] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1154–1163, 2021.

[Yu *et al.*, 2025] Qian Yu, Peng-Tao Jiang, Hao Zhang, Jinwei Chen, Bo Li, Lihe Zhang, and Huchuan Lu. High-precision dichotomous image segmentation via probing diffusion capacity. In *International Conference on Learning Representations*, 2025.

[Zhang *et al.*, 2023] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *IEEE International Conference on Computer Vision*, pages 3836–3847, 2023.

[Zhang *et al.*, 2025] Siyuan Zhang, Wei Ma, Libin Liu, Zheng Li, and Hongbin Zha. Training-free fourier phase diffusion for style transfer. In *International Joint Conference on Artificial Intelligence*, pages 2386–2394, 2025.

[Zhong and Zharkov, 2024] Yatao Zhong and Ilya Zharkov. Lightweight portrait matting via regional attention and refinement. In *IEEE Winter Conference on Applications of Computer Vision*, pages 4158–4167, 2024.

[Zhou *et al.*, 2024] Caixia Zhou, Yaping Huang, Mochu Xiang, Jiahui Ren, Haibin Ling, and Jing Zhang. Generative edge detection with stable diffusion. *arXiv preprint arXiv:2410.03080*, 2024.