This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

FlowDIS: Language-Guided Dichotomous Image Segmentation with Flow Matching

Andranik Sargsyan Shant Navasardyan
Picsart AI Research (PAIR)

https://github.com/Picsart-AI-Research/FlowDIS

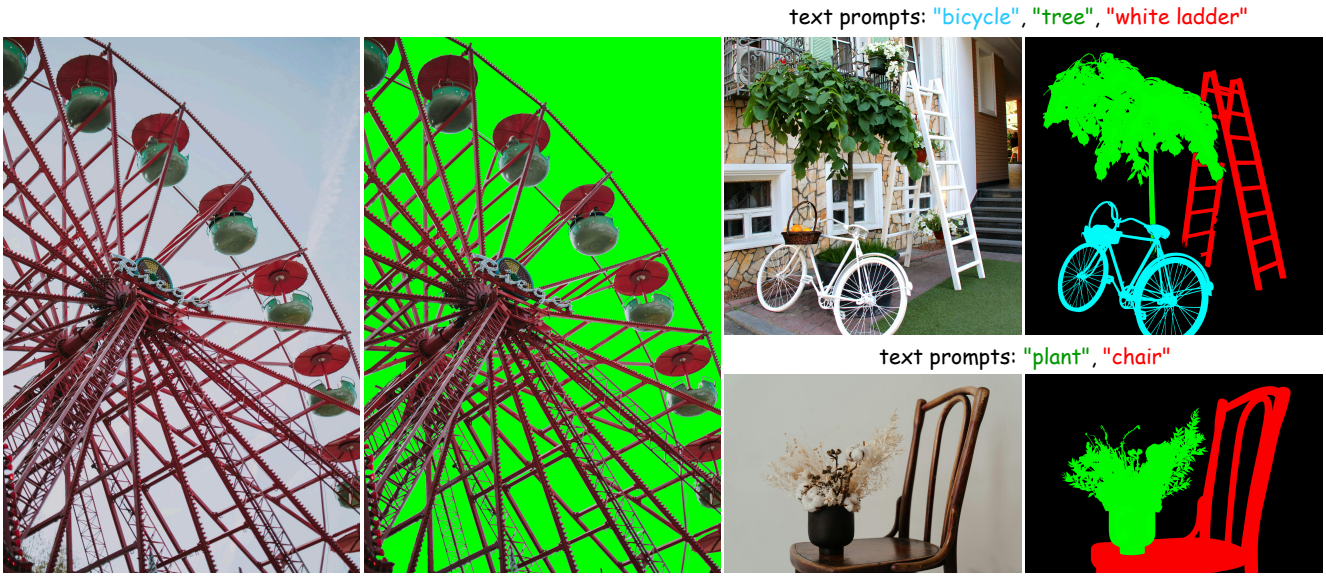


Figure 1. ***FlowDIS*** enables highly accurate foreground segmentation, *optionally* guided by a text prompt (the first example does not use text prompt). When ambiguity prevents the model from producing the desired result, the user can specify which elements to retain in the foreground. For example, in the upper-right image, if the user wants the bicycle to be segmented, they can provide the prompt "bicycle" to FlowDIS and obtain a binary mask containing only the pixels shown in cyan. Alternatively, if the user wants all three main objects to be segmented, they can prompt with "bicycle, tree, and white ladder", and FlowDIS will return a combined binary mask including all the pixels shown in cyan, green, and red.

Abstract

Accurate image segmentation is essential for modern computer vision applications such as image editing, autonomous driving, and medical image analysis. In recent years, Dichotomous Image Segmentation (DIS) has become a standard task for training and evaluating highly accurate segmentation models. Existing DIS approaches often fail to preserve fine-grained details or fully capture the semantic structure of the foreground. To address these challenges, we present ***FlowDIS****, a novel dichotomous image segmentation method built on the flow matching framework, which learns a time-dependent vector field to transport the image distribution to the corresponding mask distribution, optionally conditioned on a text prompt. Moreover, with our* ***Position-Aware Instance Pairing (PAIP)*** *training strategy, FlowDIS offers strong controllability through text prompts, enabling precise, pixel-level object segmentation. Extensive experiments demonstrate that our method significantly outperforms state-of-the-art approaches both with and without language guidance. Compared with the best prior DIS method, FlowDIS achieves a* ***5.5% higher*** *F_β^ω measure and* ***43% lower MAE*** *($\mathcal{M}$) on the DIS-TE test set. The code is available at: https://github.com/Picsart-AI-Research/FlowDIS.*

1. Introduction

Highly accurate image segmentation is crucial for modern computer vision applications, where even small errors can significantly impact downstream tasks such as image editing [16, 19], autonomous driving [13], and medical image analysis [8]. Dichotomous Image Segmentation (DIS)

[26], which involves segmenting high-precision, category-agnostic masks, has become an increasingly popular direction in the research community for developing and evaluating high-accuracy segmentation models.

Many DIS methods [9, 26, 36, 39, 40] treat segmentation as a per-pixel binary classification task and rely on pre-trained classification networks such as ResNet [6], Res2Net [4], or Swin Transformer [18] as backbones. However, since these backbones are optimized for predicting the overall class of an image, they often lack the fine-grained semantic representations required for accurate foreground segmentation, resulting in suboptimal performance on images with intricate details. Moreover, in real-world scenes containing multiple objects, classification backbones often struggle to identify and segment the correct foreground regions due to their limited ability to capture object-level semantics.

Recent progress in generative modeling [25, 29] has motivated the formulation of image segmentation within the DDPM framework [7], allowing DIS models to leverage pre-trained text-to-image (T2I) diffusion priors rather than conventional classification backbones. T2I generative models, trained on large-scale and semantically diverse datasets, provide rich representations that have proven beneficial for the downstream segmentation task. In particular, DiffDIS [37] and LawDIS [34] frame image segmentation as image-conditioned mask generation from standard Gaussian noise, building upon pre-trained Stable Diffusion [29].

However, as also noted by prior work [11, 22, 32], stochastic generative formulations are misaligned with deterministic dense prediction tasks, such as image segmentation, that require precise matches to the ground truth. This discrepancy often results in slower training convergence, requiring tens of thousands of optimization steps. Furthermore, the stochastic nature of the denoising process can blur or misplace fine boundaries, degrading the segmentation of intricate foreground structures.

To address these issues, we observe that image segmentation can be more naturally framed within the flow matching framework, enabling fully deterministic training and sampling. Flow matching offers a general framework for learning mappings between arbitrary distributions. Based on this, we propose ***FlowDIS***, which directly learns the mapping from the image distribution to the corresponding mask distribution, in contrast to the image-conditioned mask denoising (generation) paradigm of existing diffusion-based DIS methods.

Another notable property of generative models such as diffusion or flow matching models is their strong adaptability to text prompts. When adapted for the DIS task, this capability can be leveraged, as demonstrated by [34], to mitigate the ambiguity caused by multiple foreground objects within a single image. However, due to the complex multi-object nature of real-world scenes and the limited presence of multi-foreground examples in standard DIS training datasets, straightforward prompt-guided training still fails to achieve reliable language controllability, even with extensive prompting. Therefore, to further improve the language controllability, we propose the ***Position-Aware Instance Pairing (PAIP)*** strategy, which constructs mixed training examples from pairs of *(image, mask, prompt)* triplets within each training batch to provide the model with more diverse samples. Fig. 1 demonstrates the capabilities of FlowDIS for different use cases, including background removal and language-guided segmentation.

To summarize, our main contributions are the following:

- We introduce ***FlowDIS***, a highly accurate dichotomous image segmentation model leveraging the power of flow-based generative modeling.
- To enhance the language controllability of the model, we introduce a ***Position-Aware Instance Pairing (PAIP)*** strategy that selectively combines pairs of *(image, mask, prompt)* triplets within each training batch.
- FlowDIS establishes a new state-of-the-art across all test sets of DIS5K, outperforming previous methods by a significant margin. In particular, it achieves a **5.5%** higher F_β^ω score and a **43%** lower MAE ($\mathcal{M}$) on the DIS-TE test set compared to the best prior method.

2. Related Work

Research in image segmentation has evolved into several specialized subfields, each focusing on distinct types of foreground detection—such as salient [12, 30, 31, 38], camouflaged [1, 5, 21], and fine-grained [14, 35] object segmentation. To bridge these data-specific segmentation tasks, Qin et al. [26] introduced the Dichotomous Image Segmentation (DIS) task and the DIS5K dataset, which aim to provide a unified formulation and benchmark for highly accurate general foreground detection.

In recent years, DIS has attracted growing interest and motivated the development of several improved architectures. InSPyReNet [9] constructs a saliency map in an image-pyramid structure, enabling the blending of low-resolution and high-resolution scale outputs via pyramid-based image blending. FP-DIS [40] uses frequency priors to help the model capture more detailed information. Pei et al. [23] propose a dual-input network to disentangle the trunk and structure segmentation. BiRefNet [39] incorporates a bilateral reference module that draws attention to detail-rich areas during training. MVANet [36] unifies distant-view and close-up feature fusion into a single encoder–decoder architecture. PDFNet [17] introduces the depth-integrity prior to reduce false positive detections.

Recent advances in diffusion-based generative modeling [7, 29] have motivated the development of diffusion-based DIS methods [33, 34, 37]. GenPercept [33] fine-

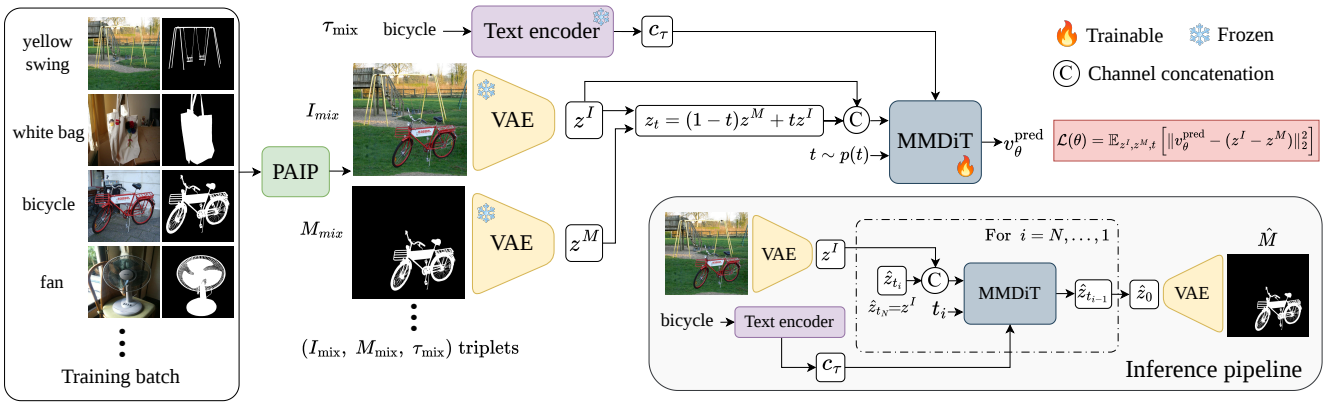


Figure 2. **Overall framework of FlowDIS.** During training, a batch of samples is passed to PAIP, which selectively combines pairs of *(image, mask, prompt)* triplets to produce a mixed batch. The mixed images and masks are encoded into the VAE latent space. For timesteps $t \sim p(t)$, intermediate latents z_t are obtained as a linear interpolation between the image and mask latents. Text prompts are encoded by the text encoder, and the resulting tokens c_τ, together with z_t, z^I and the sampled timesteps t, are fed into the MMDiT velocity prediction model. The training loss is computed as the MSE between the predicted and ground-truth velocities. During inference, the probability flow ODE is iteratively solved for $\hat{z}_0$ with the initial condition $\hat{z}_1 = z^I$, where z^I denotes the VAE encoding of the input image. The resulting latent $\hat{z}_0$ is then decoded by the VAE decoder to obtain the final mask prediction.

tunes Stable Diffusion [29] for a number of dense prediction tasks, including DIS. DiffDIS [37] uses an auxiliary edge generation task and introduces a detail-balancing attention mechanism to improve the precision of the mask prediction. LawDIS [34] integrates user controls through language guidance and window refinement. Unlike diffusion-based DIS methods, our flow matching formulation offers a more intuitive framework for image segmentation and demonstrates superior empirical performance.

3. Method

The overview of our method is shown in Fig. 2, which builds on the flow matching [15] framework. We first outline the general formulation of flow matching (Sec. 3.1), then introduce our method FlowDIS in Sec. 3.2. Finally, in Sec. 3.3 we present a Position-Aware Instance Pairing (PAIP) strategy that enhances language controllability when training data are limited.

3.1. Overview of Flow Matching

Flow matching [15] is a generative modeling approach that learns a time-dependent vector field $v_\theta(x, t)$ whose trajectories transport samples from an easy-to-sample reference distribution $p_1(x)$ to a target distribution $p_0(x)$.

Since directly learning global trajectories is intractable, flow matching instead defines a *conditional flow* between sample pairs (x_0, x_1) and trains v_θ to match the induced velocity field along the resulting path $\{x_t\}_{t\in[0,1]}$. A common choice for this conditional flow is the *linear interpolation* between the samples $x_0 \sim p_0$ and $x_1 \sim p_1$:

$$x_t = (1-t)x_0 + tx_1, \quad t \in [0,1]. \tag{1}$$

The corresponding *target velocity* along this path is

$$v(x_0, x_1) = \frac{dx_t}{dt} = x_1 - x_0. \tag{2}$$

The model $v_\theta(x, t)$ is trained to approximate this conditional vector field in expectation over p_0, p_1, and t, by minimizing the *flow matching loss*:

$$\mathcal{L}(\theta) = \mathbb{E}_{\,x_0\sim p_0, x_1\sim p_1, t}\Big[\,\|v_\theta(x_t, t) - v(x_0, x_1)\|_2^2\,\Big] \tag{3}$$

where $t \sim p(t)$ is the timestep sampling distribution during training.

Once trained, v_θ defines a continuous flow that can be integrated *backward in time* from $t = 1$ to $t = 0$ to transform samples from p_1 into realistic samples from p_0:

$$x_0 = x_1 + \int_1^0 v_\theta(x_t, t)\, dt. \tag{4}$$

In practice, this integration is performed numerically using a discretized solver such as the Euler method. Starting from a sample $x_1 \sim p_1$, the sampling process iteratively updates

$$x_{t-\Delta t} = x_t - \Delta t\, v_\theta(x_t, t), \tag{5}$$

until $t = 0$, yielding a generated sample x_0 from the target distribution.

Notably, flow matching provides a unified formulation of generative modeling in which the reference distribution p_1 can be any chosen distribution, not limited to a Gaussian prior. Diffusion models can be viewed as a special case of flow matching, where the reference dynamics are stochastic and p_1 is a standard normal distribution [15].

3.2. FlowDIS

FlowDIS leverages the power of flow matching by considering p_1 as the distribution of RGB images and p_0 as the distribution of the binary masks. Within this formulation, the velocity network is trained to transport an image $I \sim p_1$ toward its segmentation mask $M \sim p_0$, thereby learning to perform the segmentation task through deterministic flow-based generation. An overview of FlowDIS is illustrated in Fig. 2.

More precisely, given an input image $I \in \mathbb{R}^{H\times W\times 3}$ and its corresponding ground-truth mask $M \in \{0,1\}^{H\times W}$, both are encoded into a shared latent space using the encoder $\mathcal{E}(\cdot)$ of a variational autoencoder (VAE), producing the corresponding latent representations $z^I = \mathcal{E}(I)$ and $z^M = \mathcal{E}(M)$.

Following the standard optimal transport formulation of flow matching, we define the latent trajectory z_t as:

$$z_t = (1-t)z^M + tz^I, \quad t \in [0,1]. \tag{6}$$

In addition, to ensure that the velocity network always has access to the clean image signal, we condition it on the image latent z^I by concatenating it to the input[1], so the flow matching loss in Eq. (3) becomes:

$$\mathcal{L}(\theta) = \mathbb{E}_{z^I, z^M, t}\left[\|v_\theta(z_t, z^I, t) - (z^I - z^M)\|_2^2\right] \tag{7}$$

Moreover, FlowDIS extends this formulation with language-based conditioning, where the velocity network v_θ additionally takes the text embeddings c_τ of the corresponding text prompt τ as input, enabling language-guided segmentation and resulting in the final loss function:

$$\mathcal{L}(\theta) = \mathbb{E}_{z^I, z^M, t}\left[\|v_\theta(z_t, z^I, t, c_\tau) - (z^I - z^M)\|_2^2\right] \tag{8}$$

Inference of FlowDIS: Now let I be an image with one or more objects, and τ be a text prompt indicating the object the user wants to segment from I. Then FlowDIS allows segmentation by solving the probability flow ODE:

$$\frac{dz_t}{dt} = v_\theta(z_t, z^I, t, c_\tau), \quad z_1 = z^I, \tag{9}$$

where $z^I = \mathcal{E}(I)$ is the latent representation of the inference image and c_τ is the encoding of the given language prompt τ. To solve Eq. (9), we use the Euler integration method, which requires a time discretization $0 = t_0 < t_1 < \ldots < t_N = 1$, yielding the following Euler integration step:

$$z_{t_i} = z_{t_{i+1}} + v_\theta(z_{t_{i+1}}, z^I, t_{i+1}, c_\tau)\,(t_i - t_{i+1}) \tag{10}$$

for $i = N-1, \ldots, 1, 0$. After completing all steps, the final latent z_0 is decoded with the VAE decoder $\mathcal{D}(\cdot)$ to produce the predicted mask.

[1]An ablation study for this setting is included in the supplementary material.

3.3. Position-Aware Instance Pairing

Although FlowDIS inherently supports text-conditioned segmentation, we introduce a Position-Aware Instance Pairing (PAIP) strategy to further enhance its language controllability by selectively pairing *(image, mask, prompt)* triplets within each training mini-batch. This pairing encourages the model to learn from more diverse, multi-object scenes during training. The overview of PAIP is illustrated in Fig. 3.

More precisely, given a batch of triplet instances $\{(I_j, M_j, \tau_j)\}_{j=1}^n$ of size n, for each triplet (I_j, M_j, τ_j) we randomly select another triplet (I_k, M_k, τ_k) with $k \neq j$. We refer to the former as the reference sample and the latter as the pairing sample. We then combine the foreground of the pairing sample with the reference image I_j to create a new image I_{mix}, which is then paired with a foreground mask M_{mix} and prompt τ_{mix} to form the final training triplet $(I_{\text{mix}}, M_{\text{mix}}, \tau_{\text{mix}})$.

To construct the combined image I_{mix}, we first compute the minimal bounding box B_j that spans the main object indicated by the reference mask M_j. We then find the largest rectangle in the image that is adjacent to B_j—i.e., it touches B_j along one side but does not overlap (see Fig. 3 (b))—and denote it as R_j^{max}. This region defines the space where we will place the pairing foreground object from I_k. In practice, R_j^{max} is typically smaller than the bounding box B_j of the reference foreground. To compensate for this, we enlarge the available background region by padding I_j (and the corresponding mask M_j) along the side it shares with R_j^{max}. The padding amount is set equal to the length of the opposite (non-shared) side of the rectangle R_j^{max} (see Fig. 3 (c)). Reflection padding is applied to preserve visual continuity and maintain a natural appearance. This operation effectively doubles the size of the initially selected placement region R_j^{max} of I_j. We denote the padded reference image as I_j^{pad}.

To combine the object of the pairing image I_k with I_j^{pad}, we first crop its foreground region and resize it to fit within the placement area of the padded reference image I_j^{pad}, while preserving the original aspect ratio of the foreground. The resized pairing foreground is then randomly placed within this area with minimal overlap with the foreground object of I_j. Alpha blending is used to ensure seamless integration with I_j^{pad}. As a result, we get the combined image I_{mix} (see Fig. 3 (d)).

In I_{mix}, two primary objects from the reference and pairing images are present, each associated with its corresponding binary mask. Due to the padding, shifting, and resizing operations, these masks differ from the original masks M_j and M_k, and we denote the resulting ones as $\hat{M}_j$ and $\hat{M}_k$. Finally, the mask M_{mix} is randomly selected from the following generated set of options:

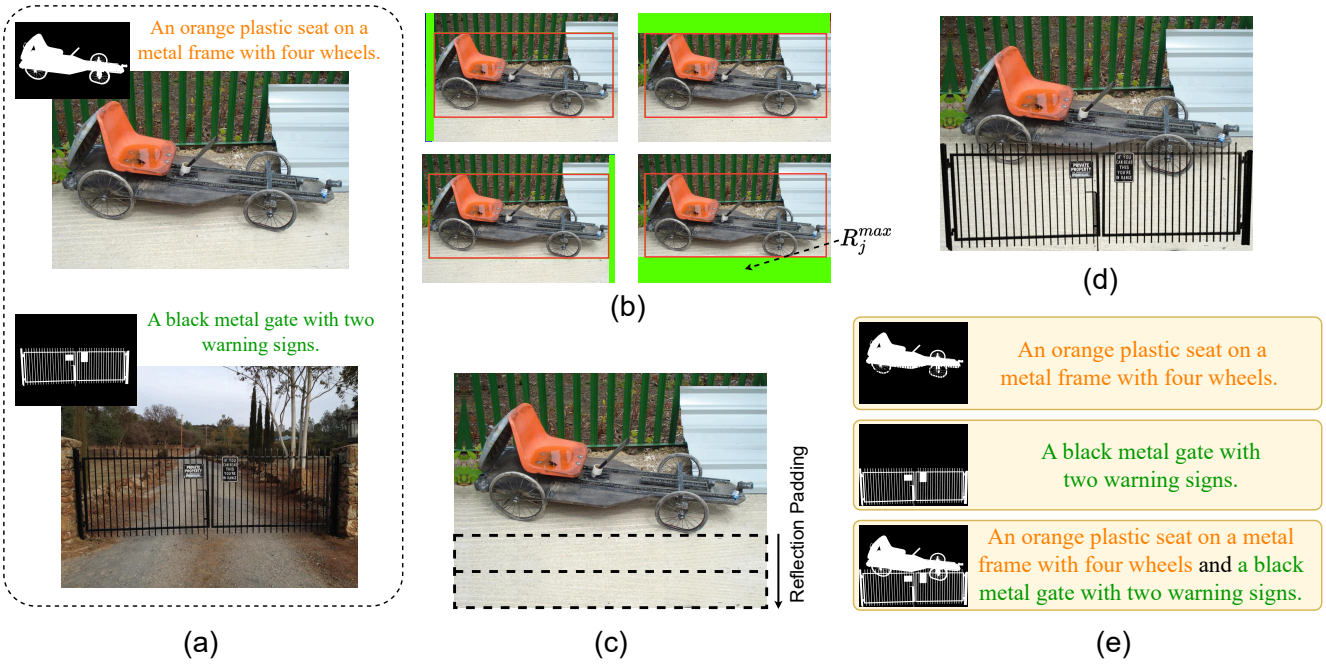


Figure 3. **Illustration of the Position-Aware Instance Pairing (PAIP) strategy. (a)** A reference sample (top triplet) is paired with another sample from the same batch (bottom triplet). **(b)** Candidate rectangular regions (in green) adjacent to the minimal bounding box (outlined in red) of the reference foreground are computed, and the one with the maximum area is selected as R_j^{max}. **(c)** The reference image is padded along the side adjacent to R_j^{max} by an amount equal to the length of its opposite side. **(d)** The pairing foreground is cropped, resized, and placed within the designated placement area. **(e)** Mask and prompt options are then constructed, from which PAIP randomly selects one for training.

$\{\hat{M}_j$ AND $(\hat{M}_k)^c$, $\hat{M}_k$, $\hat{M}_j$ OR $\hat{M}_k\}$, where $(\hat{M}_k)^c$ denotes the complement of $\hat{M}_k$, i.e. $(\hat{M}_k)^c = 1 - \hat{M}_k$, AND denotes the pixel-wise multiplication, and OR denotes the pixel-wise maximum. We then select the corresponding textual description τ_{mix} from $\{\tau_j, \tau_k, \text{“}\tau_j \text{ and } \tau_k\text{”}\}$ according to the chosen M_{mix} as shown in Fig. 3 (e).

4. Experiments

4.1. Implementation Details

As the base flow matching MMDiT model, we adopt FLUX.1-Schnell [10], initialized with its pre-trained weights. To incorporate an additional image latent condition, we extend the input channels of the first linear layer in the transformer, initializing the new weights to zeros. For text guidance, CLIP [27] and T5 [28] are used as text encoders to obtain the token sequence c_τ. During training, the timestep distribution $p(t)$ is a $\text{Beta}(2.5, 1)$ distribution, which biases the training toward larger t values, where prediction is more challenging. The models are trained with a batch size of 32 for 10,000 iterations ($\approx$ 1.8 days) on 8 $\times$ NVIDIA A100 GPUs. For optimization, we use the AdamW optimizer with an initial learning rate of 5×10^{-5}, which is halved at steps 512, 2048, 4096, and 8192.

For inference, we adopt a non-uniform schedule de-

Algorithm 1 FlowDIS Inference

Require: v-pred. model v_θ, VAE $(\mathcal{E}, \mathcal{D})$, image I, text condition c_τ, steps $N \in \mathbb{N}$, $\alpha, \beta \in \mathbb{R}^+$

$q \leftarrow \text{linspace}(0, 1, N+1)$
$t_i \leftarrow F^{-1}_{\text{Beta}}(q_i; \alpha, \beta), \quad i = 0, \ldots, N$ # Beta schedule
$\hat{z}_{t_N} = z^I \leftarrow \mathcal{E}(I)$ # Encode image into latent space
for $n = N-1, \ldots, 0$ **do**
  $\hat{z}_{t_n} \leftarrow \hat{z}_{t_{n+1}} + v_\theta(\hat{z}_{t_{n+1}}, z^I, t_{n+1}, c_\tau)(t_n - t_{n+1})$
end for
$\hat{M}_{\text{rgb}} \leftarrow \mathcal{D}(\hat{z}_0)$ # Decode latent to RGB mask
$\hat{M}_{i,j} \leftarrow \frac{1}{3}\sum_{k=1}^{3}(\hat{M}_{\text{rgb}})_{i,j,k}$ # Convert to grayscale
$\hat{M} \leftarrow \text{clip}(\hat{M}, 0, 1)$
return $\hat{M}$ # Final predicted mask

rived from the Beta cumulative distribution function (CDF). Given N total timesteps, we first define an equidistant grid $q \in [0, 1]$ of length $N+1$, which is then mapped using the inverse Beta CDF:

$$t_i = F^{-1}_{\text{Beta}}(q_i; \alpha, \beta), \quad i = 0, \ldots, N, \tag{11}$$

where $F^{-1}_{\text{Beta}}(\cdot; \alpha, \beta)$ denotes the inverse CDF with shape parameters $\alpha, \beta \in \mathbb{R}^+$. This schedule enables denser sampling near the start or end of the flow trajectory depending

Methods	DIS-TE1 (500 images)					DIS-TE2 (500 images)					DIS-TE3 (500 images)				
	$F_\beta^\omega \uparrow$	$F_\beta^{mx} \uparrow$	$\mathcal{M} \downarrow$	$\mathcal{S}_\alpha \uparrow$	$E_\phi^{mn} \uparrow$	$F_\beta^\omega \uparrow$	$F_\beta^{mx} \uparrow$	$\mathcal{M} \downarrow$	$\mathcal{S}_\alpha \uparrow$	$E_\phi^{mn} \uparrow$	$F_\beta^\omega \uparrow$	$F_\beta^{mx} \uparrow$	$\mathcal{M} \downarrow$	$\mathcal{S}_\alpha \uparrow$	$E_\phi^{mn} \uparrow$
IS-Net$_{22}$ [26]	0.662	0.740	0.074	0.787	0.820	0.728	0.799	0.070	0.823	0.858	0.758	0.830	0.064	0.836	0.883
InSPyReNet$_{22}$ [9]	0.788	0.845	0.043	0.873	0.894	0.846	0.894	0.036	0.905	0.928	0.871	0.919	0.034	0.918	0.943
FP-DIS$_{23}$ [40]	0.713	0.784	0.060	0.821	0.860	0.767	0.827	0.059	0.845	0.893	0.811	0.868	0.049	0.871	0.922
UDUN$_{23}$ [23]	0.720	0.784	0.059	0.817	0.864	0.768	0.829	0.058	0.843	0.886	0.809	0.865	0.050	0.865	0.917
BiRefNet$_{24}$ [39]	0.820	0.860	0.037	0.885	0.912	0.858	0.893	0.036	0.900	0.931	0.894	0.925	0.028	0.919	0.957
GenPercept$_{24}$ [33]	0.794	0.844	0.038	0.871	0.909	0.827	0.875	0.040	0.887	0.925	0.840	0.890	0.039	0.893	0.939
MVANet$_{24}$ [36]	0.825	0.873	0.037	0.887	0.916	0.879	0.916	0.030	0.918	0.943	0.891	0.929	0.030	0.923	0.952
DiffDIS$_{25}$ [37]	0.820	0.895	0.035	0.900	0.905	0.859	0.923	0.032	0.922	0.927	0.877	0.940	0.032	0.929	0.936
PDFNet$_{25}$ [17]	0.845	0.888	0.031	0.887	0.916	0.884	0.919	0.029	0.921	0.946	0.888	0.929	0.029	0.923	0.953
LawDIS$_{25}$ [34]	0.866	0.899	0.029	0.906	0.934	0.888	0.921	0.030	0.920	0.947	0.899	0.929	0.028	0.924	0.955
Ours (1-step)	0.939	**0.961**	**0.012**	**0.953**	0.975	0.944	**0.965**	0.013	**0.958**	0.976	0.939	0.962	0.016	**0.954**	0.973
Ours (2-step)	**0.942**	**0.961**	**0.012**	**0.953**	**0.976**	**0.947**	**0.965**	**0.012**	**0.958**	**0.978**	**0.942**	**0.963**	**0.015**	**0.954**	**0.975**

Methods	DIS-TE4 (500 images)					DIS-TE (1-4) (2,000 images)					DIS-VD (470 images)				
	$F_\beta^\omega \uparrow$	$F_\beta^{mx} \uparrow$	$\mathcal{M} \downarrow$	$\mathcal{S}_\alpha \uparrow$	$E_\phi^{mn} \uparrow$	$F_\beta^\omega \uparrow$	$F_\beta^{mx} \uparrow$	$\mathcal{M} \downarrow$	$\mathcal{S}_\alpha \uparrow$	$E_\phi^{mn} \uparrow$	$F_\beta^\omega \uparrow$	$F_\beta^{mx} \uparrow$	$\mathcal{M} \downarrow$	$\mathcal{S}_\alpha \uparrow$	$E_\phi^{mn} \uparrow$
IS-Net$_{22}$ [26]	0.753	0.827	0.072	0.830	0.870	0.726	0.799	0.070	0.819	0.858	0.717	0.791	0.074	0.813	0.856
InSPyReNet$_{22}$ [9]	0.848	0.905	0.042	0.905	0.928	0.838	0.891	0.039	0.900	0.923	0.834	0.889	0.042	0.900	0.922
FP-DIS$_{23}$ [40]	0.788	0.846	0.061	0.852	0.906	0.770	0.830	0.057	0.847	0.895	0.763	0.823	0.062	0.843	0.891
UDUN$_{23}$ [23]	0.792	0.846	0.059	0.849	0.901	0.772	0.831	0.057	0.844	0.892	0.763	0.823	0.059	0.838	0.892
BiRefNet$_{24}$ [39]	0.865	0.904	0.039	0.900	0.941	0.858	0.896	0.035	0.901	0.934	0.855	0.891	0.038	0.898	0.932
GenPercept$_{24}$ [33]	0.801	0.861	0.055	0.869	0.918	0.816	0.868	0.043	0.880	0.923	0.815	0.865	0.043	0.881	0.922
MVANet$_{24}$ [36]	0.865	0.912	0.038	0.908	0.939	0.862	0.907	0.034	0.909	0.938	0.863	0.904	0.034	0.908	0.936
DiffDIS$_{25}$ [37]	0.835	0.915	0.045	0.911	0.910	0.848	0.918	0.036	0.916	0.919	0.844	0.915	0.037	0.917	0.917
PDFNet$_{25}$ [17]	0.867	0.910	0.038	0.909	0.941	0.874	0.913	0.031	0.913	0.943	0.873	0.913	0.030	0.915	0.944
LawDIS$_{25}$ [34]	0.884	0.922	0.034	0.915	0.952	0.884	0.918	0.030	0.916	0.947	0.884	0.917	0.030	0.917	0.949
Ours (1-step)	0.912	0.945	0.026	0.938	0.961	0.933	0.958	0.017	**0.951**	0.971	0.933	0.957	0.015	0.952	0.972
Ours (2-step)	**0.919**	**0.946**	**0.024**	**0.939**	**0.964**	**0.938**	**0.959**	**0.016**	**0.951**	**0.973**	**0.938**	**0.958**	**0.014**	**0.953**	**0.974**

Table 1. Quantitative comparison with 10 representative methods on the DIS5K dataset. All results are evaluated at 1024×1024 px input resolution. $\downarrow$ indicates lower is better, while $\uparrow$ means higher is better. The best and second-best results are highlighted in **bold** and underlined, respectively.

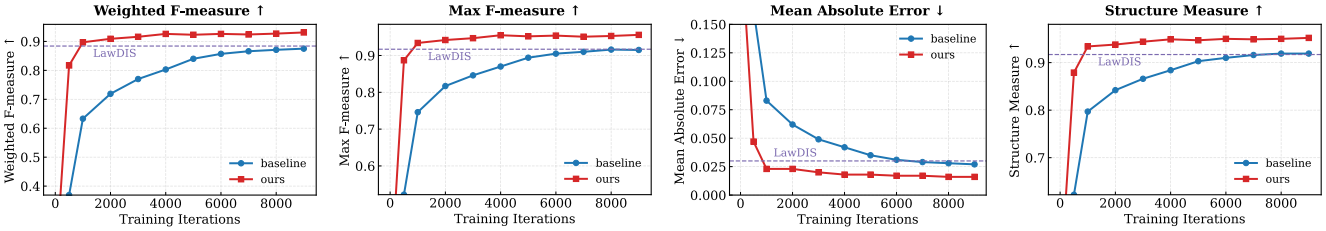


Figure 4. Metrics evaluated at different training iteration checkpoints for our method vs the baseline (image-conditioned mask generation from noise). Dashed lines show the metric levels of fully trained LawDIS [34]. Our FlowDIS converges faster than the baseline denoising approach and surpasses state-of-the-art LawDIS after only 1K training iterations, whereas LawDIS was trained for 36K iterations.

on (α, β), providing finer control over integration dynamics. We use $\alpha = 2.5$ and $\beta = 1.0$, consistent with training. Our full inference pipeline is summarized in Algorithm 1.

4.2. Experimental Setup

Dataset: We conduct our experiments on the DIS5K [26] dataset, which consists of 5,470 high-resolution image-mask pairs across 225 categories. It is further divided into DIS-TR (3,000 images), DIS-VD (470 images), and DIS-TE (2,000 images). All training experiments are conducted on DIS-TR, while DIS-VD and DIS-TE are exclusively used for testing. For language guidance, we generate captions τ using a vision-language model. We discuss this process in detail in the supplementary material.

DIS-TE is further divided into DIS-TE1, DIS-TE2, DIS-TE3, and DIS-TE4 subsets, each containing 500 images, where the numbers 1–4 denote increasing levels of foreground complexity. We test our approach on all subsets separately, as well as on the combined test set DIS-TE (1-4), to compare FlowDIS across all complexity levels.

Evaluation metrics: We use widely adopted metrics: the weighted F-measure ($F_\beta^\omega \uparrow$) [20], max F-measure ($F_\beta^{mx} \uparrow$) [24], mean absolute error ($\mathcal{M} \downarrow$) [24], Structure-measure ($S_\alpha \uparrow$) [2] and E-measure ($E_\phi^{mn} \uparrow$) [3].

4.3. Quantitative and Qualitative Analysis

Quantitative comparison. We compare our FlowDIS with the following state-of-the-art DIS methods: IS-Net [26], In-

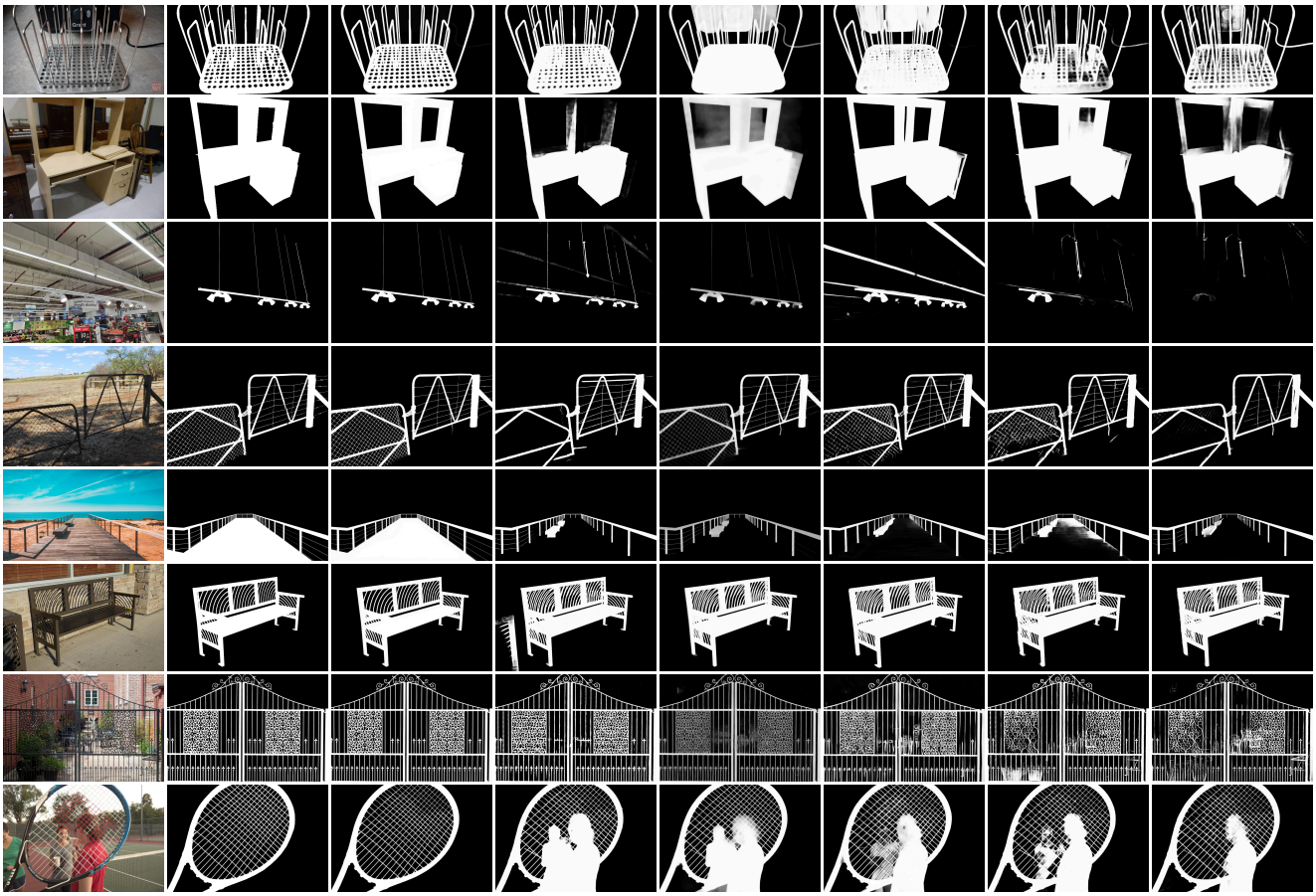


Figure 5. Qualitative comparison with state-of-the-art DIS methods. Our approach generates more detailed and more semantically accurate masks.

SPyReNet [9], FP-DIS [40], UDUN [23], BiRefNet [39], GenPercept [33], MVANet [36], DiffDIS [37], PDFNet [17] and LawDIS [34]. For a fair comparison, all methods are evaluated at an input resolution of 1024×1024 px. As shown in Tab. 1, our method already achieves state-of-the-art results across all test sets with 1-step inference, while using 2-step inference yields even better performance. For example, on the full DIS test set, 1-step FlowDIS yields an $\approx 5.5\%$ improvement in F_β^ω and an $\approx 43\%$ reduction in $\mathcal{M}$ over the runner-up language-guided model LawDIS [34].

Qualitative comparison. Fig. 5 presents a qualitative comparison between FlowDIS and other state-of-the-art methods. As shown in rows 1, 4, and 7, our method captures fine-grained visual details more effectively, while in rows 2, 3, and 5, it demonstrates superior semantic understanding of the scene.

In Fig. 6, we compare the language controllability of our method with the state-of-the-art language-guided approach, LawDIS [34]. As shown, our method exhibits stronger controllability while simultaneously producing more accurate results. For instance, in the second row, our method correctly isolates the black sedan from a background containing two other cars, whereas LawDIS produces nearly identical outputs regardless of the input prompt. Notably, as can be seen from our qualitative results, while the introduced PAIP strategy constructs samples with a pair of foregrounds, FlowDIS generalizes to more complex scenes by successfully selecting the requested object. More qualitative results can be found in the supplementary material.

4.4. Ablation Studies

We conduct ablation studies on the DIS-VD subset using fixed 2-step inference unless specified otherwise.

Effectiveness of our deterministic flow matching formulation: We evaluate the effectiveness of the flow matching setup described in Sec. 3.2. Specifically, we compare our approach (deterministic FM), which uses the image latent as z_1, against a variant where z_1 is sampled from standard

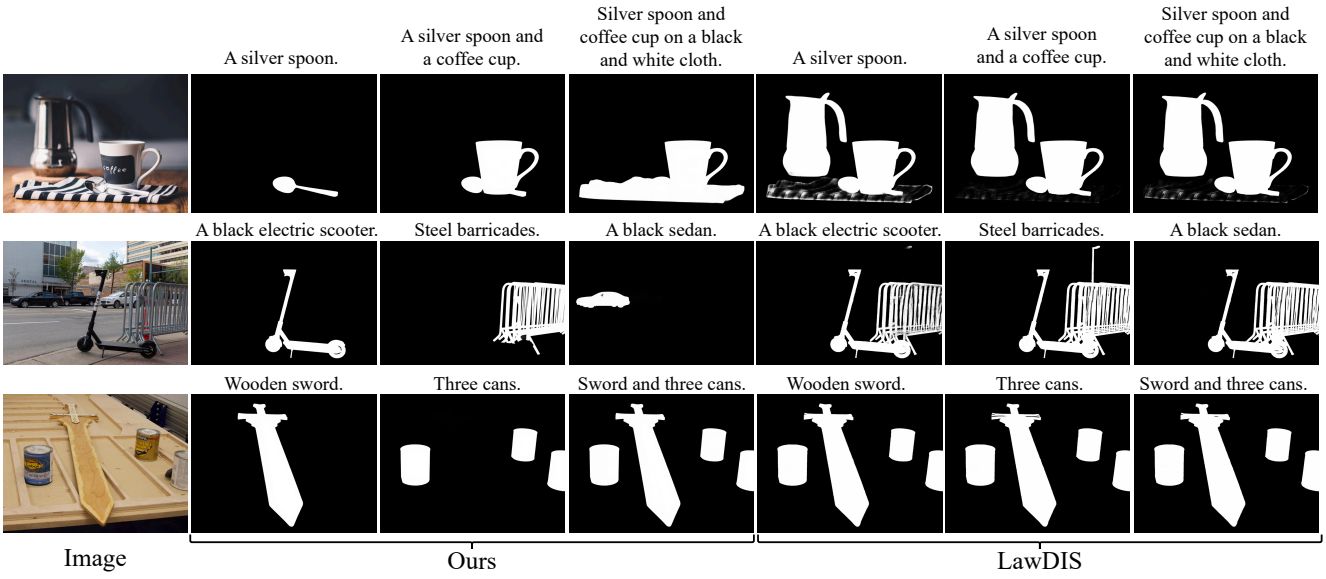


Figure 6. Comparison of language controllability between our method and LawDIS [34]. Each output is generated using the corresponding text prompt shown above.

Gaussian noise[2] during training and inference (denoising FM). As can be seen from Tab. 2, our deterministic formulation significantly outperforms the denoising-based variant.

Ablation settings	$F_\beta^\omega \uparrow$	$F_\beta^{mx} \uparrow$	$\mathcal{M} \downarrow$	$\mathcal{S}_\alpha \uparrow$	$E_\phi^{mn} \uparrow$
denoising FM	0.883	0.916	0.025	0.920	0.957
deterministic FM	**0.938**	**0.958**	**0.014**	**0.953**	**0.974**

Table 2. Ablation study on flow matching (FM) settings.

We also compare the training convergence of the baseline denoising-based flow matching approach with our FlowDIS formulation by evaluating the models at different training iterations and plotting their performance. As shown in Fig. 4, our formulation converges significantly faster than the baseline. Moreover, our method requires only 1K training iterations to surpass the second-best model, LawDIS [34], while LawDIS was trained for 36K iterations.

Effectiveness of language guidance: We train and evaluate FlowDIS with and without language prompts. As shown in Tab. 3, incorporating language guidance significantly improves performance by providing semantic cues that help the model resolve ambiguous inputs.

Ablation settings	$F_\beta^\omega \uparrow$	$F_\beta^{mx} \uparrow$	$\mathcal{M} \downarrow$	$\mathcal{S}_\alpha \uparrow$	$E_\phi^{mn} \uparrow$
w/o language guidance	0.901	0.926	0.027	0.929	0.951
w/ language guidance	**0.937**	**0.956**	**0.015**	**0.952**	**0.975**

Table 3. Ablation study on language guidance.

Effectiveness of PAIP: Since PAIP is intended to improve language controllability and the DIS-VD validation set contains mainly single objects, we construct a new benchmark by applying PAIP to the DIS-VD subset. The resulting test set, DIS-VD-Complex, maintains the same number of samples as DIS-VD but introduces more complex visual compositions, which can be challenging for models with weaker language-following capabilities. As shown in Tab. 4, PAIP significantly improves the results on the complex-scene test set DIS-VD-Complex, while preserving performance on the simple-scene set DIS-VD.

	DIS-VD-Complex			DIS-VD		
Ablation settings	$F_\beta^{mx} \uparrow$	$\mathcal{M} \downarrow$	$\mathcal{S}_\alpha \uparrow$	$F_\beta^{mx} \uparrow$	$\mathcal{M} \downarrow$	$\mathcal{S}_\alpha \uparrow$
FlowDIS **w/o** PAIP	0.783	0.063	0.831	0.956	0.015	0.952
FlowDIS **w/** PAIP	**0.960**	**0.014**	**0.955**	**0.958**	**0.014**	**0.953**

Table 4. Ablation study on PAIP.

Additional quantitative and qualitative ablation results on PAIP are provided in the supplementary material.

5. Conclusion

We presented a novel, language-guided dichotomous image segmentation approach, ***FlowDIS***, which reformulates the segmentation task within the flow matching framework. Our formulation connects the predictive image segmentation task with generative modeling, while preserving the deterministic nature of segmentation. To further improve language controllability, we introduced a ***Position-Aware Instance Pairing (PAIP)*** strategy, which constructs pairwise foreground compositions within each training batch while selecting the guidance prompt and the target mask. Extensive quantitative and qualitative experiments show that our approach surpasses existing state-of-the-art methods.

[2]We still provide z^I to the MMDiT v_θ via concatenation to the input.

References

[1] Zhongxi Chen, Ke Sun, and Xianming Lin. Camodiffusion: Camouflaged object detection via conditional diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1272–1280, 2024. 2

[2] Ming-Ming Cheng and Deng-Ping Fan. Structure-measure: A new way to evaluate foreground maps. *International Journal of Computer Vision (IJCV)*, 129(9):2622–2638, 2021. 6

[3] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. In *International Joint Conferences on Artificial Intelligence*, 2018. 6

[4] Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip Torr. Res2Net: A new multi-scale backbone architecture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(2):652–662, 2019. 2

[5] Chunming He, Kai Li, Yachao Zhang, Longxiang Tang, Yulun Zhang, Zhenhua Guo, and Xiu Li. Camouflaged object detection with feature decomposition and edge reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22046–22055, 2023. 2

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 2

[7] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020. 2

[8] Ge-Peng Ji, Jingyi Liu, Peng Xu, Nick Barnes, Fahad Shahbaz Khan, Salman Khan, and Deng-Ping Fan. Frontiers in intelligent colonoscopy. *Machine Intelligence Research*, 23 (1):70–114, 2026. 1

[9] Taehun Kim, Kunhee Kim, Joonyeong Lee, Dongmin Cha, Jiho Lee, and Daijin Kim. Revisiting image pyramid structure for high resolution salient object detection. In *Proceedings of the Asian Conference on Computer Vision*, pages 108–124, 2022. 2, 6, 7

[10] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 5

[11] Hsin-Ying Lee, Hung-Yu Tseng, Hsin-Ying Lee, and Ming-Hsuan Yang. Exploiting diffusion prior for generalizable dense prediction. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7861–7871, Los Alamitos, CA, USA, 2024. IEEE Computer Society. 2

[12] Guanbin Li and Yizhou Yu. Deep contrast learning for salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 478–487, 2016. 2

[13] Ruibo Li, Hanyu Shi, Zhe Wang, and Guosheng Lin. Weakly and self-supervised class-agnostic motion prediction for autonomous driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 1

[14] Jun Hao Liew, Scott Cohen, Brian Price, Long Mai, and Jiashi Feng. Deep interactive thin object selection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 305–314, 2021. 2

[15] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. 3

[16] Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025. 1

[17] Xianjie Liu, Keren Fu, and Qijun Zhao. High-precision dichotomous image segmentation via depth integrity-prior and fine-grained patch strategy. *arXiv preprint arXiv:2503.06100*, 2025. 2, 6, 7

[18] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 2

[19] Guangben Lu, Yuzhen Du, Yizhe Tang, Zhimin Sun, Ran Yi, Yifan Qi, Tianyi Wang, Lizhuang Ma, and Fangyuan Zou. Pinco: Position-induced consistent adapter for diffusion transformer in foreground-conditioned inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15266–15276, 2025. 1

[20] Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. How to Evaluate Foreground Maps. In *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, Los Alamitos, CA, USA, 2014. IEEE Computer Society. 6

[21] Haiyang Mei, Ge-Peng Ji, Ziqi Wei, Xin Yang, Xiaopeng Wei, and Deng-Ping Fan. Camouflaged object segmentation with distraction mining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8772–8781, 2021. 2

[22] Ziqi Pang, Xin Xu, and Yu-Xiong Wang. Aligning generative denoising with discriminative objectives unleashes diffusion for visual perception. *arXiv preprint arXiv:2504.11457*, 2025. 2

[23] Jialun Pei, Zhangjun Zhou, Yueming Jin, He Tang, and Pheng-Ann Heng. Unite-divide-unite: Joint boosting trunk and structure for high-accuracy dichotomous image segmentation. In *Proceedings of the 31st ACM International Conference on Multimedia*, page 2139–2147, New York, NY, USA, 2023. Association for Computing Machinery. 2, 6, 7

[24] Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung. Saliency filters: Contrast based filtering for salient region detection. In *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 733–740, Los Alamitos, CA, USA, 2012. IEEE Computer Society. 6

[25] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[26] Xuebin Qin, Hang Dai, Xiaobin Hu, Deng-Ping Fan, Ling Shao, and Luc Van Gool. Highly accurate dichotomous im-

age segmentation. In *European Conference on Computer Vision*, 2022. 2, 6

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 5

[28] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. 5

[29] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, Los Alamitos, CA, USA, 2022. IEEE Computer Society. 2, 3

[30] Lv Tang, Bo Li, Yijie Zhong, Shouhong Ding, and Mofei Song. Disentangled high quality salient object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3580–3590, 2021. 2

[31] Wenguan Wang, Shuyang Zhao, Jianbing Shen, Steven CH Hoi, and Ali Borji. Salient object detection with pyramid attention and salient edges. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1448–1457, 2019. 2

[32] Changliang Xia, Chengyou Jia, Minnan Luo, Zhuohang Dang, Xin Shen, and Bowen Ping. D^3-predictor: Noise-free deterministic diffusion for dense prediction. *arXiv preprint arXiv:2512.07062*, 2025. 2

[33] Guangkai Xu, Yongtao Ge, Mingyu Liu, Chengxiang Fan, Kangyang Xie, Zhiyue Zhao, Hao Chen, and Chunhua Shen. What matters when repurposing diffusion models for general dense perception tasks? In *The Thirteenth International Conference on Learning Representations*, 2025. 2, 6, 7

[34] Xinyu Yan, Meijun Sun, Ge-Peng Ji, Fahad Shahbaz Khan, Salman Khan, and Deng-Ping Fan. LawDIS: Language-window-based controllable dichotomous image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23902–23911, 2025. 2, 3, 6, 7, 8

[35] Chenglin Yang, Yilin Wang, Jianming Zhang, He Zhang, Zhe Lin, and Alan Yuille. Meticulous object segmentation. *arXiv preprint arXiv:2012.07181*, 2020. 2

[36] Qian Yu, Xiaoqi Zhao, Youwei Pang, Lihe Zhang, and Huchuan Lu. Multi-view aggregation network for dichotomous image segmentation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3921–3930, Los Alamitos, CA, USA, 2024. IEEE Computer Society. 2, 6, 7

[37] 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 *The Thirteenth International Conference on Learning Representations*, 2025. 2, 3, 6, 7

[38] Yi Zeng, Pingping Zhang, Jianming Zhang, Zhe Lin, and Huchuan Lu. Towards high-resolution salient object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7234–7243, 2019. 2

[39] Peng Zheng, Dehong Gao, Deng-Ping Fan, Li Liu, Jorma Laaksonen, Wanli Ouyang, and Nicu Sebe. Bilateral reference for high-resolution dichotomous image segmentation. *CAAI Artificial Intelligence Research*, 3:9150038, 2024. 2, 6, 7

[40] Yan Zhou, Bo Dong, Yuanfeng Wu, Wentao Zhu, Geng Chen, and Yanning Zhang. Dichotomous image segmentation with frequency priors. In *International Joint Conferences on Artificial Intelligence*, 2023. 2, 6, 7