LawDIS: Language-Window-based Controllable Dichotomous Image Segmentation

Xinyu Yan[1,2,6] Meijun Sun[1,2] Ge-Peng Ji[3]
Fahad Shahbaz Khan[6] Salman Khan[6] Deng-Ping Fan[4,5*]
[1] Tianjin University [2] Tianjin Key Laboratory of Machine Learning [3] Australian National University
[4] Nankai Institute of Advanced Research (SHENZHEN FUTIAN) [5] Nankai University [6] MBZUAI

Abstract

We present LawDIS, a language-window-based controllable dichotomous image segmentation (DIS) framework that produces high-quality object masks. Our framework recasts DIS as an image-conditioned mask generation task within a latent diffusion model, enabling seamless integration of user controls. LawDIS is enhanced with macro-to-micro control modes. Specifically, in macro mode, we introduce a language-controlled segmentation strategy (LS) to generate an initial mask based on user-provided language prompts. In micro mode, a window-controlled refinement strategy (WR) allows flexible refinement of user-defined regions (i.e., size-adjustable windows) within the initial mask. Coordinated by a mode switcher, these modes can operate independently or jointly, making the framework well-suited for high-accuracy, personalised applications. Extensive experiments on the DIS5K benchmark reveal that our LawDIS significantly outperforms 11 cutting-edge methods across all metrics. Notably, compared to the second-best model MVANet, we achieve F_β^ω gains of 4.6% with both the LS and WR strategies and 3.6% gains with only the LS strategy on DIS-TE. Codes will be made available at https://github.com/XinyuYanTJU/LawDIS.

1. Introduction

With the popularity of high-quality camera equipment, segmentation task in computer vision has evolved from rough localisation [23, 37] to high-precision delineation [47]. As a result, the dichotomous image segmentation (DIS) task [32] – focused on segmenting highly accurate foreground object(s) – has garnered significant attention due to its broad range of applications, *e.g.*, 3D reconstruction [22, 38], image editing [10, 21], augmented reality [31, 44], and medical image segmentation [12, 14].

Compared to general segmentation tasks [19, 20, 36], the DIS task presents greater challenges in two key aspects.

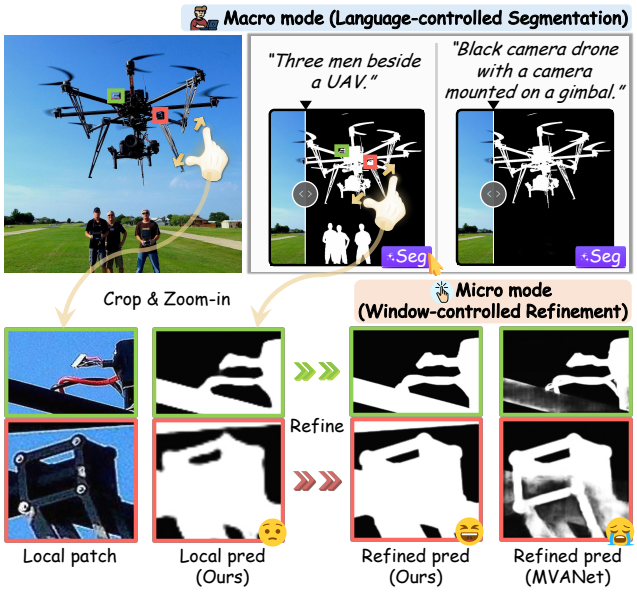

Figure 1. Illustration of the proposed macro and micro controls for the DIS task. The macro mode enables users to segment objects with customised language prompts, while micro mode supports post-refinement on user-defined windows at any scale. After refinement, our results become more precise, whereas the runner-up model, MVANet [45], lacks adaptability to the cropped local patches, resulting to inferior predictions.

First, the widely recognised DIS benchmark [32] covers more than 200 object categories and diverse visual traits, such as salient [8, 51] and camouflaged [6, 9, 13] objects. This diversity requires models to exhibit strong holistic understanding to accurately identify foreground objects across varied scenarios. Additionally, this task emphasizes highly precise object delineation in high-resolution images, capturing even the internal details of objects. Therefore, models should exhibit fine-grained feature extraction capabilities to effectively segment complex structures and shapes.

Current DIS methods [16, 26, 32, 45, 49, 50] often adopt a discriminative learning paradigm (*i.e.*, per-pixel classifi-

*Corresponding author: Deng-Ping Fan (dengpfan@gmail.com)

cation) that depends on their intrinsic learning ability to derive object semantics. This paradigm would struggle in real-world applications, particularly when addressing personalised needs. First, when an image contains multiple foreground entities, it is semantically ambiguous which object(s) should be targeted. Second, to better capture geometric details of high-resolution objects, most methods [16, 26, 42] incorporate an extra high-resolution data flow, often by downsampling a full-size image to a 1024^2px input. This straightforward way compensates for fine-grained details of a segmented object, while infinitely scaling up the input size is computationally impractical. Recent methods [45, 49] address this by splitting the full-size image into a series of patches, equivalently enlarging geometric details with fewer pixel losses. However, as these methods are trained on patches with predefined size, they lack adaptability to variable patch sizes. As shown in Fig. 1, the advanced model, MVANet [45], fails when given a local patch of input size that differs from those used during training.

To alleviate the above issues, we present a language-windows-based controllable DIS framework, LawDIS, designed to meet user-personalized requirements. Our framework recasts the DIS as an image-conditioned mask generation task within a latent diffusion model [34], enabling the seamless integration of various user controls. Furthermore, we enhance this generative framework with macro-to-micro control modes. In macro mode, we propose a language-controlled segmentation strategy (LS) that generates an initial mask based on a user-provided language prompt, as shown in the upper part of Fig. 1. In micro mode, the window-controlled refinement strategy (WR) refines user-specified regions at variable scales based on the initial mask, as illustrated on the lower part of Fig. 1. These two modes can operate independently or jointly through a mode switcher, facilitating adaptive optimisation during training and mutual refinement during inference. Extensive experiments on the DIS5K benchmark [32] reveals that LawDIS significantly surpasses 11 state-of-the-art (SOTA) methods across all metrics. Specifically, compared to the second-best model, MVANet, we achieve an improvement of 4.6% in F_β^ω when employing both LS and WR strategies, and an improvement of 3.6% with the only LS strategy.

Our main contributions are summarized as follows:

- We present LawDIS, a language-window-based controllable framework that reformulates DIS as a image-conditioned mask generation problem, enabling seamless user-controlled integration for producing high-quality object masks.
- This framework is enhanced with macro-to-micro control modes. In macro mode, we introduce a language-controlled segmentation strategy (LS) to generate an initial mask based on the user-provided language prompt. In micro mode, we design a window-controlled refinement strategy (WR) to refine user-specified regions (*i.e.*, size-adjustable windows) within this initial mask.
- These two modes can operate independently or jointly via a mode switcher, resulting in SOTA performance across all metrics on the DIS5K benchmark.

2. Related Works

Dichotomous image segmentation. Qin *et al*. [32] introduce the DIS task for accurately segmenting objects with varying structural complexities in high-resolution images, regardless of their characteristics, which has garnered significant interest from the research community. Following the advent of the fully convolutional network [24], many models formulate this task within a discriminative framework, treating it as a per-pixel classification problem. Early solutions [26, 29, 30, 32, 39, 50] rely on convolutional designs, employing strategies such as intermediate supervision [32], frequency priors [50], and unite-divide-unite [26] to improve performance. Recently, visual transformers [5] have gained favour due to their powerful capability to model long-range dependencies, resulting in impressive results. However, due to the absence of convolutional local inductive bias, their ability to capture local structures remains relatively weak. To overcome this limitation, InSPyReNet [16], BiRefNet [49], and MVANet [45] introduce additional images or patches of varying resolutions as input during the training or inference stages, enriching the capture of detailed information to some extent.

The aforementioned DIS methods, which rely on discriminative learning paradigms, primarily focus on balancing and integrating global and local information but tend to overlook two key challenges: flexible semantic control across different scenarios and local window refinement for unavoidable blurry segmentation. Different from these methods, the proposed LawDIS reformulates the DIS task within a generative-based model, leveraging the encyclopedic visual-language understanding and the advantages of denoising mechanisms to achieve language-controlled segmentation and window-controlled refinement.

Diffusion models for high-resolution segmentation. Currently, a trend in the computer vision community adapting latent diffusion model [34] to high-resolution segmentation tasks [46]. Considering the challenges associated with high-resolution data acquisition, a reasonable approach [28] is to leverage the powerful generative capabilities of stable diffusion to synthesize data, thereby enhancing the performance of existing high-resolution segmentation methods. In addition, Wang *et al*. [40] propose a refinement model to enhance the quality of object masks generated by different segmentation models using the diffusion process. Later, a two-stage latent diffusion approach [41] has employed for portrait matting. GenPercept [43] transforms the generative model into a deterministic one-step fine-tuning paradigm

through a customized decoder and conducts extensive experiments on a wide range of fundamental visual dense perception tasks, including high-resolution segmentation tasks such as DIS and portrait matting. Distinct from prior methods, we extend a single stable diffusion to both macro and micro modes using a switcher. This not only enables both high-resolution image segmentation and result refinement within the same model, but also allows user control over both language and window aspects.

3. Methodology

We propose a user-controlled framework via reformulating DIS task within an image-conditioned mask generation paradigm. Our framework supports two levels of control. Specifically, the macro-mode allows users to specify and segment the objects from the high-resolution image, guided by custom language prompts. Meanwhile, the micro-mode functions as a general post-refinement tool to improve the accuracy of segmentation masks, particularly in areas with intricate structures. For better integrability, we consolidate macro- and micro-modes within one diffusion model using a mode switcher, yielding better geometric representations at varied scales. Next, we introduce the preliminaries to build the generative paradigm (Sec. 3.1) and two key processes (Sec. 3.2 & Sec. 3.3) to repurpose it for the DIS task.

3.1. Generative Formulation for DIS

We redefine the DIS task involving conditional denoising diffusion, focusing on modelling the conditional probability distribution $D(s \mid x)$ for segmentation masks $s \in \mathbb{R}^{W\times H}$. The condition $x \in \mathbb{R}^{W\times H\times 3}$ is specified by the corresponding RGB image awaiting segmentation. Utilizing the framework established by a pre-trained Stable Diffusion v2 model [34], our LawDIS efficiently executes the diffusion process within a lower-dimensional latent space. Initially, a variational autoencoder with an encoder $\phi(\cdot)$ and a decoder $\varphi(\cdot)$ is used to transform data between segmentation mask and latent space, such that $\mathbf{z}^{(s)} = \phi(s)$ and $s \approx \varphi(\mathbf{z}^{(s)})$. Similarly, x is transposed to latent representation as $\mathbf{z}^{(x)} = \phi(x)$, serving as a condition for generation.

Second, stable diffusion sets up a *forward* nosing process and a *reversal* denoising process within the latent space using a U-Net. In *forward* process, starting from $\mathbf{z}_0^{(s)} := \mathbf{z}^{(s)}$, Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$ is gradually added at each level $t \in \{1, \ldots, T\}$ to construct a discrete Markov chain $\{\mathbf{z}_0^{(s)}, \mathbf{z}_1^{(s)}, \ldots, \mathbf{z}_T^{(s)}\}$.

In *reversal* process, the U-Net model $f_\theta(\cdot)$ with learned parameters θ, gradually predicts the noise $\boldsymbol{\epsilon}$ added to the noisy sample $\mathbf{z}_t$ at each step t. θ is updated as follows: a data pair (s, x) is sampled from the training set and converted to $(\mathbf{z}^{(s)}, \mathbf{z}^{(x)})$, noise $\boldsymbol{\epsilon}$ is added to $\mathbf{z}^{(s)}$ at a random time step t, and the estimated noise $\hat{\boldsymbol{\epsilon}}$ is computed as

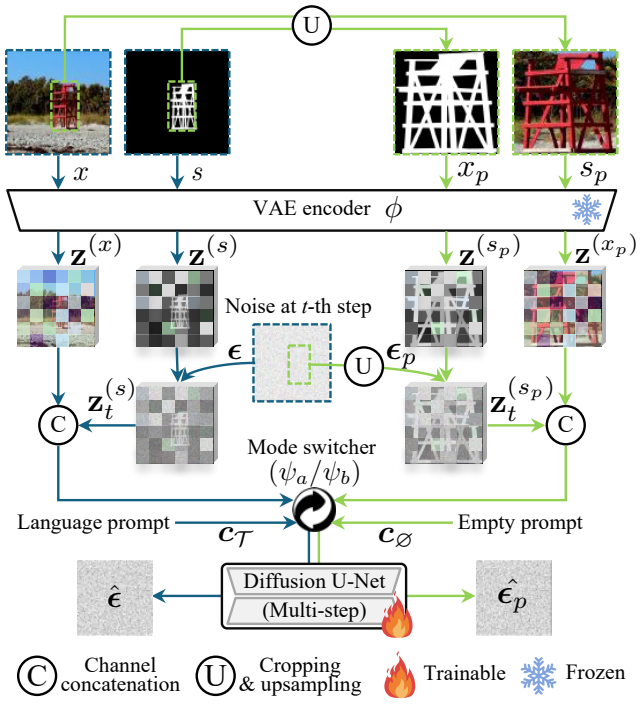

Figure 2. An overview of the U-Net training protocol in LawDIS. Starting from the pretrained stable diffusion, we introduce a mode switcher to extend it to both macro and micro modes for joint training, where ψ_a activates macro mode and ψ_b activates micro mode. The entire image x and its segmentation map s serve as inputs for the macro mode, while its patch image x_p and patch map s_p serve as inputs for the micro mode. Both sets of input are fed into the VAE encoder, converting them into latent space. Paired with language prompt $\boldsymbol{c}_\mathcal{T}$ and empty prompt $\boldsymbol{c}_\varnothing$ respectively, these latent representations are then fed into the U-Net, resulting in $\hat{\boldsymbol{\epsilon}}$ and $\hat{\boldsymbol{\epsilon}}_p$, which are sent for optimization of the standard diffusion objective relative to segmentation latent codes.

$f_\theta(\mathbf{z}_t^{(s)}, \mathbf{z}^{(x)}, t)$. The objective is to minimize:

$$\mathbb{E}_{\epsilon\sim\mathcal{N}(0,1),t,\mathbf{z}_0^{(s)},\mathbf{z}^{(x)}}[\|\boldsymbol{\epsilon} - f_\theta(\mathbf{z}_t^{(s)}, \mathbf{z}^{(x)}, t)\|_2^2]. \quad (1)$$

During inference, noise $\hat{\boldsymbol{\epsilon}}$ can be progressively predicted starting from a normally distributed variable $\mathbf{z}_T^{(s)} \in \mathcal{N}(\mathbf{0}, \mathbf{1})$, and then gradually denoised under the guidance of denoising schedulers (such as DDPM [11], DDIM [35], *etc.*) to obtain $\mathbf{z}_0^{(s)}$. Subsequently, the estimated clean latent representation $\mathbf{z}_0^{(s)}$ is reconstructed through the decoder φ to get the mask prediction $\hat{s} = \varphi(\mathbf{z}_0^{(s)}) \sim D(s \mid x)$.

3.2. Joint Training with Two Modes

Mode switcher. To enable LawDIS to perform different functions in two modes, we introduce a mode switcher ψ to stable diffusion, which is represented as a one-dimensional vector encoded by the positional encoding and then added

with the time embeddings of diffusion model. ψ is set to either ψ_a or ψ_b, which activates the macro mode or the micro mode, respectively. The two modes are designed to mutually enhance each other during training and enable seamless switching during inference.

Macro mode. When ψ_a is activated, LawDIS switches to the macro mode, where a language-controlled segmentation strategy is employed to segment objects with the guidance of the user-provided prompt. During training, the model receives the full image x, segmentation mask s and its corresponding user prompts $\mathcal{T}$. These prompts are generated by VLM [1, 2]. See supp. for more details. As described in Sec. 3.1, x and s are encoded into latent space, which are then passed into the diffusion process to learn to predict the added Gaussian noise $\boldsymbol{\epsilon}$. As per [34], we use CLIP [33] to encode prompts $\mathcal{T}$ into a control embedding $\boldsymbol{c}_{\mathcal{T}}$, which is integrated into the diffusion U-Net via cross-attention mechanism. The loss function for the macro mode is:

$$\mathcal{L}_{macro} = \|\boldsymbol{\epsilon} - f_\theta(\mathbf{z}_t^{(s)}, \mathbf{z}^{(x)}, \boldsymbol{c}_{\mathcal{T}}, t, \psi_a)\|_2^2. \quad (2)$$

Micro mode. When ψ_b is selected, LawDIS activates its micro mode, employing a window-controlled refinement strategy to precisely delineate details within user-specified windows. During training, we select the minimum enclosing rectangle of the foreground object in the segmentation mask s as the local window, rather than using random windows. This ensures that the foreground object is fully within the window, thus preventing missing it or introducing ambiguous semantics. Based on this window selection criteria, we crop the corresponding regions from the image x and the segmentation mask s, resulting in local patch x_p and local mask s_p. Accordingly, the noise $\boldsymbol{\epsilon}$ is cropped to $\boldsymbol{\epsilon}_p$. To avoid mismatches between the local patch/mask and user prompt, an empty prompt $\varnothing$ is used. These inputs are processed through the UNet to facilitate learning of local noise prediction. The loss function for micro-aware process is:

$$\mathcal{L}_{micro} = \|\boldsymbol{\epsilon}_p - f_\theta(\mathbf{z}_t^{(s_p)}, \mathbf{z}^{(x_p)}, \boldsymbol{c}_{\varnothing}, t, \psi_b)\|_2^2. \quad (3)$$

Joint training recipe. To achieve collaborative enhancement between the two modes, we jointly train the U-Net $f_\theta(\cdot)$ with both modes, as shown in Fig. 2. The naive diffusion loss Equ.(1) is reformulated as the sum of the losses from each mode: $\mathcal{L}_u = \mathcal{L}_{macro} + \mathcal{L}_{micro}$.

After training the U-Net, we fine-tune the VAE decoder φ initially designed for RGB image reconstruction, to adapt it for the DIS task. First, to preserve the effectiveness of the previous training results, we freeze both the encoder ϕ and the U-Net during this process, while only fine-tuning the decoder φ. Second, we make simple structural adjustments to the decoder φ by adding shortcut connections between the encoder and decoder. Moreover, the output layer of the decoder reduces the channels from 3 to 1, with its weight

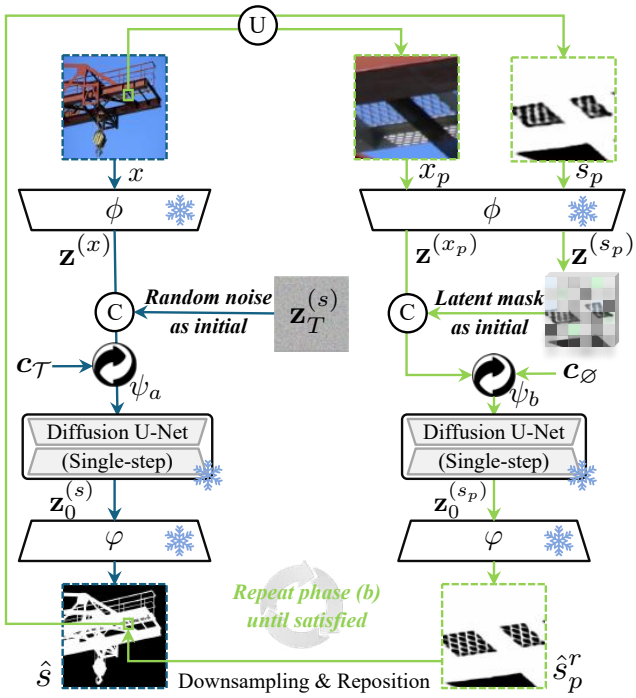

Figure 3. Overview of the inference protocol, which consists of two steps. The first step is language-controlled segmentation, where LawDIS switches to the macro mode to generate an initial segmentation result based on the language prompt. If the user is not satisfied with the details, the second step, window-controlled refinement, is executed to refine details within a controllable window of variable resolution. The refined local patches are then used to replace their corresponding regions in the initial result. The second step can be repeated indefinitely until a satisfactory segmentation result is achieved.

tensor initialized by averaging across the channels. Next, we randomly input pairs (image x, text prompt $\mathcal{T}$, ψ_a) or (local patch x_p, empty text prompt $\varnothing$, ψ_b) with a normally distributed noise $\mathbf{z}_T^{(s)} \in \mathcal{N}(\mathbf{0}, \mathbf{1})$ into the model. These input pass sequentially through the VAE encoder, U-Net, denoising scheduler, and VAE decoder to obtain the predicted segmentation mask $\hat{s}$ or local mask $\hat{s}_p$. The process is supervised using the annotated mask s or s_p, with the structure loss function defined as follows:

$$\mathcal{L}_d = \begin{cases} \mathcal{L}_{wbce}(\hat{s}, s) + \mathcal{L}_{wiou}(\hat{s}, s) & \text{if } \psi = \psi_a, \\ \mathcal{L}_{wbce}(\hat{s}_p, s_p) + \mathcal{L}_{wiou}(\hat{s}_p, s_p) & \text{if } \psi = \psi_b. \end{cases} \quad (4)$$

where $\mathcal{L}_{wbce}(\cdot)$ and $\mathcal{L}_{wiou}(\cdot)$ are defined as [45, 50]. It is important to note that, in order to obtain $\hat{s}$ or $\hat{s}_p$, the model needs to performs T denoising steps to generate the clean latent features $\mathbf{z}_0^{(s)}$ or $\mathbf{z}_0^{(s_p)}$ from random noise. However, using denoising schedulers like DDIM [35], which typically

require 50 steps to generate segmentation images, proves impractical in realistic settings [18]. Therefore, we introduce the trajectory consistency distillation (TCD) [48] as an out-of-the-box denoising scheduler to simplify the sampling process to a single step. It not only enables feasible fine-tuning of the VAE decoder φ avoiding out-of-memory, but also enhances the inference efficiency.

3.3. Two-stage Inference

Fig. 3 provides an overview of the inference process, which consists of two steps. The first step is language-guided segmentation, where LawDIS switches to macro mode to generate segmentation results based on language prompts. The second step serves as an optional refinement stage, invoked only when user-driven adjustments are required. During this process, the micro mode is selected, allowing the user to refine details within a controllable window of variable resolution. It can be repeated indefinitely until a satisfactory segmentation result is achieved. The details of both steps are provided below.

Language-controlled segmentation. The switcher is modulated by ψ_a which activates the macro mode. Given the input image x, we use the VAE encoder ϕ to transform it into the latent space $\mathbf{z}^{(x)} = \phi(x)$. Then, we initialize the segmentation mask latent as standard Gaussian noise $\mathbf{z}_T^{(s)} \in \mathcal{N}(\mathbf{0}, \mathbf{1})$. The language prompts used to control the segmented objects $\mathcal{T}$ can be generated by VLM or customized by users. Next, they are fed into the denoising U-Net together to predict noise. A single-step denoising is performed with the TCD scheduler [48] to obtain the clean latent features $\mathbf{z}_0^{(s)}$. Finally, by decoding the features $\mathbf{z}_0^{(s)}$ with the fine-tuned VAE decoder, we obtain the language-controlled segmentation map $\hat{s} = \varphi(\mathbf{z}_0^{(s)})$.

Window-controlled refinement. The switcher is set to ψ_b which activates the micro mode. A key challenge is ensuring reliable refinement when feeding local patches cropped from arbitrary windows into the network, as they lack context from the full image. To address this, we propose using local patches from the global segmentation result instead of noise as the starting point for diffusion, indirectly transferring contextual information between the two modes.

Specifically, the user can click on any area in the initial segmentation map to select the unsatisfactory region as the window for refinement. For each window for refinement, two patches are cropped: the local patch x_p, cropped from the input image x, and the initial local mask $\hat{s}_p$, cropped from the language-controlled segmentation map $\hat{s}$ obtained in the first step. These patches are upsampled to the input size of the model. Next, we utilize the encoder ϕ to obtain the latent features of the local patch $\mathbf{z}^{(x_p)} = \phi(x_p)$ and the local mask $\mathbf{z}^{(\hat{s}_p)} = \phi(\hat{s}_p)$, respectively. We change the language prompt $\mathcal{T}$ to empty $\varnothing$ as a condition. The latent features of the initial local mask $\mathbf{z}^{(\hat{s}_p)}$ are combined with those of the local patch $\mathbf{z}^{(x_p)}$ and fed into the denoising U-Net to predict noise. After denoising with the TCD denoising scheduler and decoding with the VAE decoder, a refined local mask $\hat{s}_p^r$ with clearer details is obtained. Finally, $\hat{s}_p^r$ replaces the original content at its designated location in the language-controlled segmentation map $\hat{s}$. This process allows multiple windows to be refined simultaneously. When there is an overlapping region between multiple windows, the value from the window with a larger proportion of the overlapping area is chosen as the final optimization result.

4. Experiments

4.1. Experimental setups

Dataset. We perform all experiments on the DIS5K benchmark [32], which contains 5,470 high-resolution image-mask pairs across 225 semantic categories. This benchmark is divided into DIS-TR (3,000 images), DIS-VD (470 images), and DIS-TE (2,000 images). We conduct all training on DIS-TR and evaluate all models on DIS-VD and DIS-TE. The DIS-TE has four subsets, from DIS-TE1 to DIS-TE4, each containing 500 samples, to represent increasing levels of shape complexities.

Evaluation. For a comprehensive comparison, we employ five widely used pixel-wise metrics to assess the ability of the models, including the weighted F-measure (F_β^ω) [25], the maximum F-measure (F_β^{mx}) [27], the structure measure ($\mathcal{S}_\alpha$) [4], the mean enhanced-alignment measure (E_ϕ^{mn}) [7], and the mean absolute error (M) [27].

Implementation details. We implement LawDIS using PyTorch library, accelerated by a single NVIDIA A100-40GB GPU. To repurpose Stable Diffusion v2 [34] for conditional mask generation, we duplicate the diffusion U-Net's input layer to match the concatenated image feature and noised mask feature. We initialise the duplicated layers via copying the weight tensor and halve its values to avoid inflation of activation size [15]. We employ the DDPM noise scheduler [11] with 1000 steps in the diffusion U-Net training stage, while the VAE decoder is trained using a single-step TCD scheduler [48] due to computational constraints. Both stages use a batch size of 32 and Adam optimizer with a 3e-5 learning rate, while diffusion U-Net is fine-tuned for 30K iterations and VAE decoder for 6K iterations. To increase visual diversity, a random horizontal flipped augmentation and an annealed multi-resolution noise strategy are applied. For better inference efficiency, we use the TCD noise scheduler in both macro and micro modes. All inputs, whether full-size images or window-sized regions, are uniformly resized to 1024^2px for training and inference.

4.2. Comparison with SOTA Methods

Quantitative comparison. As presented in Tab. 1, we establish a DIS benchmark with 11 well-known task-related methods, including BASNet [29], U^{2}Net [30], HRNet [39],

Methods	DIS-TE1 (500 images)					DIS-TE2 (500 images)					DIS-TE3 (500 images)				
	$F_\beta^\omega \uparrow$	$F_\beta^{mx} \uparrow$	$\mathcal{M} \downarrow$	$\mathcal{S}_\alpha \uparrow$	$E_\phi^{mn} \uparrow$	$F_\beta^\omega \uparrow$	$F_\beta^{mx} \uparrow$	$\mathcal{M} \downarrow$	$\mathcal{S}_\alpha \uparrow$	$E_\phi^{mn} \uparrow$	$F_\beta^\omega \uparrow$	$F_\beta^{mx} \uparrow$	$\mathcal{M} \downarrow$	$\mathcal{S}_\alpha \uparrow$	$E_\phi^{mn} \uparrow$
BASNet_{19} [29]	0.577	0.663	0.105	0.741	0.756	0.653	0.738	0.096	0.781	0.808	0.714	0.790	0.080	0.816	0.848
$\text{U}^2\text{Net}_{20}$ [30]	0.601	0.701	0.085	0.762	0.783	0.676	0.768	0.083	0.798	0.825	0.721	0.813	0.073	0.823	0.856
HRNet_{20} [39]	0.579	0.668	0.088	0.742	0.797	0.664	0.747	0.087	0.784	0.840	0.700	0.784	0.080	0.805	0.869
PGNet_{22} [42]	0.680	0.754	0.067	0.800	0.848	0.743	0.807	0.065	0.833	0.880	0.785	0.843	0.056	0.844	0.911
IS-Net_{22} [32]	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} [16]	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} [50]	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} [26]	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} [49]	0.819	0.860	0.037	0.885	0.911	0.857	0.894	0.036	0.900	0.930	0.893	0.925	0.028	0.919	0.955
GenPercept_{24} [43]	0.794	0.844	0.038	0.871	0.909	0.828	0.875	0.040	0.887	0.925	0.840	0.890	0.039	0.893	0.939
MVANet_{24} [45]	0.820	0.870	0.037	0.885	0.914	0.875	0.915	0.030	0.917	0.943	0.888	0.929	0.029	0.923	0.953
Ours-S	**0.886**	**0.917**	**0.025**	**0.917**	**0.946**	**0.906**	**0.934**	**0.024**	**0.932**	**0.958**	**0.908**	**0.937**	**0.025**	**0.931**	**0.960**
Ours-R	**0.890**	**0.919**	**0.024**	**0.917**	**0.948**	**0.914**	**0.936**	**0.022**	**0.932**	**0.961**	**0.919**	**0.942**	**0.022**	**0.932**	**0.964**

Methods	DIS-TE4 (500 images)					DIS-TE (1-4) (2,000 images)					DIS-VD (470 images)				
	$F_\beta^\omega \uparrow$	$F_\beta^{mx} \uparrow$	$\mathcal{M} \downarrow$	$\mathcal{S}_\alpha \uparrow$	$E_\phi^{mn} \uparrow$	$F_\beta^\omega \uparrow$	$F_\beta^{mx} \uparrow$	$\mathcal{M} \downarrow$	$\mathcal{S}_\alpha \uparrow$	$E_\phi^{mn} \uparrow$	$F_\beta^\omega \uparrow$	$F_\beta^{mx} \uparrow$	$\mathcal{M} \downarrow$	$\mathcal{S}_\alpha \uparrow$	$E_\phi^{mn} \uparrow$
BASNet_{19} [29]	0.713	0.785	0.087	0.806	0.844	0.664	0.744	0.092	0.786	0.814	0.656	0.737	0.094	0.781	0.809
$\text{U}^2\text{Net}_{20}$ [30]	0.707	0.800	0.085	0.814	0.837	0.676	0.771	0.082	0.799	0.825	0.656	0.753	0.089	0.785	0.809
HRNet_{20} [39]	0.687	0.772	0.092	0.792	0.854	0.658	0.743	0.087	0.781	0.840	0.641	0.726	0.095	0.767	0.824
PGNet_{22} [42]	0.774	0.831	0.065	0.841	0.899	0.746	0.809	0.063	0.830	0.885	0.733	0.798	0.067	0.824	0.879
IS-Net_{22} [32]	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} [16]	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} [50]	0.788	0.846	0.061	0.852	0.906	0.770	0.831	0.047	0.847	0.895	0.763	0.823	0.062	0.843	0.891
UDUN_{23} [26]	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} [49]	0.864	0.904	0.039	0.900	0.939	0.858	0.896	0.035	0.901	0.934	0.854	0.891	0.038	0.898	0.931
GenPercept_{24} [43]	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} [45]	0.866	0.913	0.038	0.910	0.940	0.862	0.907	0.034	0.909	0.938	0.861	0.904	0.035	0.909	0.937
Ours-S	**0.890**	**0.926**	**0.032**	**0.920**	**0.955**	**0.898**	**0.929**	**0.027**	**0.925**	**0.955**	**0.894**	**0.925**	**0.026**	**0.924**	**0.955**
Ours-R	**0.910**	**0.932**	**0.026**	**0.922**	**0.964**	**0.908**	**0.932**	**0.024**	**0.926**	**0.959**	**0.905**	**0.929**	**0.023**	**0.925**	**0.959**

Table 1. Quantitative comparison of DIS5K with 11 representative methods. ↓ represents the lower value is better, while ↑ represents the higher value is better. The best and the second-best results are highlighted in **red** and **blue**.

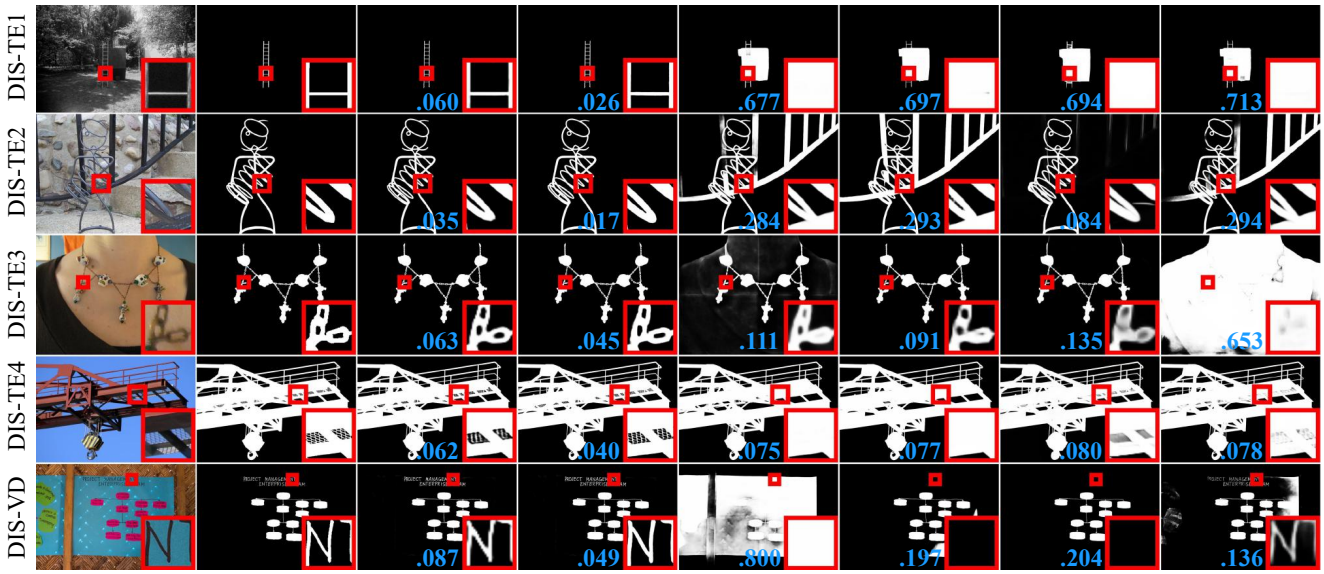

Figure 4. Qualitative comparison of our model with four leading models. Local masks are evaluated with $\mathcal{M}$ score for clarity.

PGNet [42], IS-Net [32], InSPyReNet [16], FP-DIS [50], UDUN [26], BiRefNet [49], GenPercept [43], and MVANet [45]. With the same setup of input size (1024^2px), our basic configuration with language controls (Ours-S), has outperformed the competing methods [16, 26, 32, 39, 42, 43, 45, 49, 50] in all metrics across all test sets. These results underscore the efficacy of LawDIS with user prompts in addressing the challenging cases of DIS, especially for diverse categories of targets. For example, Ours-S surpasses the runner-up model MVANet [45] by 6.6% in F_β^ω on DIS-TE1.

Ablation settings	$F_\beta^{mx} \uparrow$	$\mathcal{M} \downarrow$	$\mathcal{S}_\alpha \uparrow$	$E_\phi^{mn} \uparrow$
baseline [34]	0.904	0.047	0.904	0.916
w/o micro-level training	0.912	0.037	0.909	0.943
w/o fine-tuning VAE decoder	0.919	0.040	0.915	0.933
Ours-S	**0.926**	**0.032**	**0.920**	**0.955**

Table 2. Ablation study for the baseline and general settings.

Ablation settings	$F_\beta^{mx} \uparrow$	$\mathcal{M} \downarrow$	$\mathcal{S}_\alpha \uparrow$	$E_\phi^{mn} \uparrow$
w/o user prompt (train & test)	0.912	0.036	0.908	0.944
w/o user prompt (test only)	0.915	0.036	0.909	0.947
***w/* user prompt (train & test)**	**0.926**	**0.032**	**0.920**	**0.955**

Table 3. Ablation study on macro controls for LawDIS.

In micro mode, Ours-R refines the initial predictions from the macro mode (Ours-S), leading to a further 2.0% increase in F_β^ω on DIS-TE4. Here, to simulate subjectivity of user interaction, window candidates are chosen along the object's contours of the GT. See supp. for more details. By integrating dual control mechanisms into LawDIS, we achieve a gain of 7.0% in F_β^ω on DIS-TE1 compared to MVANet.

Qualitative comparison. Fig. 4 provides a qualitative comparison between our approach and the most competitive existing DIS models. At the macro level, we achieve more complete segmentation of the target regions, while at the micro level, it demonstrates superior precision in handling complex structures and intricate details. For example, as shown in the 5th row, the edges of the subtitles appear sharper, and the $\mathcal{M}$ score for local patches shows a drop of 3.8% from Ours-S (0.087) to Ours-R (0.049).

4.3. Diagnostic Study

Next, we assess the impact of key components through a series of ablation studies conducted on DIS-TE4 subset.

Baseline and general settings. To reveal the true potential of stable diffusion [34] in the DIS task, we initialise a baseline model by modifying LawDIS with three changes: omitting the mode switcher, training the diffusion U-Net without user prompts, and not fine-tuning VAE decoder. As presented in the first row of Tab. 2, the baseline variant does not achieve superior performance. To assess the impact of the joint training strategy, we remove the mode switcher and train LawDIS exclusively using language controls. This variant (second row of Tab. 2) exhibits performance drops across all metrics relative to Ours-S, demonstrating the dual-mode synergy's role in providing scalable geometric representations of for varied input sizes. In addition, we create another variant (third row of Tab. 2) where the VAE decoder is not fine-tuned, and its output is averaged across channels to produce a single-channel mask prediction. The inferior performance demonstrates that fine-tuning the VAE decoder is essential for high-resolution segmentation, as it enables the decoding process to complement the denoised mask features with fine-grained details.

Effectiveness of macro controls. We design experiments under two settings: in the first row of Tab. 3, the user prompts are set to empty during the training and testing

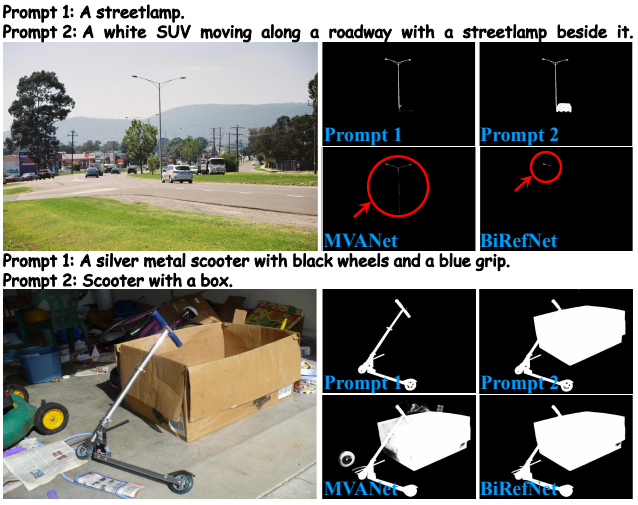

Figure 5. Qualitative predictions under different macro controls.

Ablation settings	$F_\beta^\omega \uparrow$	$\mathcal{M} \downarrow$	$BIoU^m \uparrow$	$HCE_\gamma \downarrow$
basic setting (*i.e.*, Ours-S)	0.890	0.032	0.795	2481
init. from Gaussian noise	-4.7%	+1.9%	-7.1%	-863
auto windows selection	+1.7%	-0.5%	+2.9%	-767
semi-auto windows selection	**+2.0%**	**-0.6%**	**+3.2%**	**-871**

Table 4. Ablation study on micro controls for LawDIS.

phases; in the second row, user prompts are omitted only during the testing phase. Compared to them, our default setting, which uses user prompts in both phases, achieves superior results. As shown in Fig. 5, our model showcases the ability to flexibly segment various target objects based on customized language prompts. In contrast, other methods [45, 49], which lack the ability to process language prompts, yield only fixed results through memory patterns during training, further emphasizing the effectiveness of LawDIS.

Effectiveness of micro controls. To better unveil performance change on fine structures, we introduce two extra metrics, human correction effort (HCE_γ) [32] and boundary intersection-over-union ($BIoU^m$) [3]. In the second row of Tab. 4, we replace the patch mask latent with Gaussian noise as input, resulting in a performance drop from 0.890 to 0.843 in F_β^ω. This suggests that initiating the diffusion process using the segmentation results can help the model achieve more refined masks. In addition, the third row of Tab. 4 presents a fully automated window selection process, where windows are chosen around object edges in the initial predictions from Ours-S without user intervention. The result demonstrates that the WR strategy effectively improves segmentation performance, even without user intervention in the window selection.

4.4. Discussion

Can the proposed WR strategy function as a general post-refinement tool? To reveal its compatibility, we apply the WR strategy to refine the initial masks predicted by various out-of-the-box DIS methods [16, 26, 32, 45, 49]. As indicated in Tab. 5, we improve the performance of each

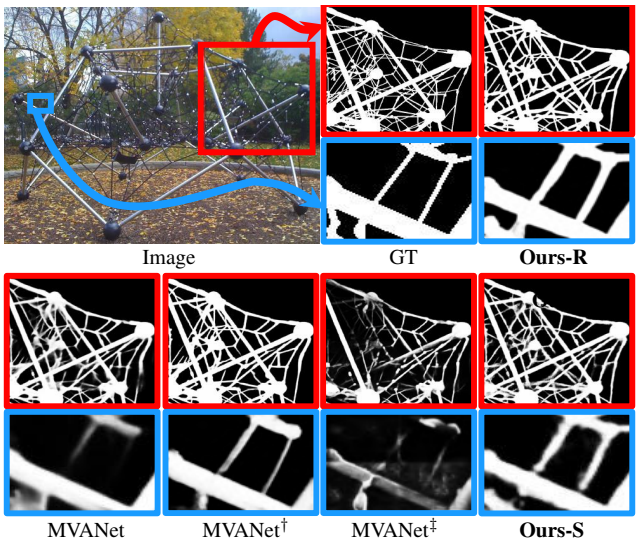

Figure 6. Qualitative comparison of refinement performance of WR strategy and segmentation performance on local patches of MVANet [45]. † represents the refined segmentation results by applying WR strategy, while ‡ represents the segmentation result of the local patch obtained by other methods (*e.g.*, MVANet).

Methods	$F_\beta^\omega \uparrow$	$\mathcal{M} \downarrow$	$\mathcal{S}_\alpha \uparrow$	$E_\phi^{mn} \uparrow$
IS-Net† [32]	0.753 +6.4%	0.072 -1.6%	0.830 +2.7%	0.870 +4.2%
InSPyReNet† [16]	0.848 +4.2%	0.042 -1.1%	0.905 +0.8%	0.928 +2.6%
UDUN† [26]	0.792 +3.9%	0.059 -1.0%	0.849 +2.0%	0.901 +2.2%
BiRefNet† [49]	0.864 +2.5%	0.039 -0.6%	0.900 +0.7%	0.939 +1.3%
MVANet† [45]	0.866 +3.2%	0.038 -0.9%	0.910 +0.6%	0.940 +1.8%

Table 5. Using WR strategy as a post-refinement tool to enhance current DIS methods. The performance gains are underlined.

Methods	$F_\beta^\omega \uparrow$	$\mathcal{M} \downarrow$	$\mathcal{S}_\alpha \uparrow$	$E_\phi^{mn} \uparrow$
IS-Net‡ [32]	0.753 -3.6%	0.072 +1.0%	0.830 -4.5%	0.870 -4.5%
InSPyReNet‡ [16]	0.848 -0.8%	0.042 +0.7%	0.905 -2.0%	0.928 -0.5%
UDUN‡ [26]	0.792 -3.8%	0.059 +1.6%	0.849 -3.6%	0.901 -2.8%
BiRefNet‡ [49]	0.864 -3.9%	0.039 +1.6%	0.900 -4.3%	0.939 -3.1%
MVANet‡ [45]	0.866 -0.2%	0.038 +0.6%	0.910 -1.4%	0.940 -0.7%
Ours‡ (Ours-R)	0.890 **+ 2.0%**	0.032 **- 0.6%**	0.920 **+ 0.2%**	0.955 **+ 0.9%**

Table 6. Quantitative comparison of the segmentation performance on local patches.

model in the DIS-TE4 subset to varying degrees. For example, the predictions of IS-Net [32] and MVANet [45] achieve 6.4% and 3.2% increases in F_β^ω, respectively. Fig. 6 compares the zoom-in region from the full-image segmentation strategy (MVANet) to the refined patch mask obtained via our WR strategy (MVANet†). This demonstrate the feasibility of WR strategy as a post-refinement tool to enhance existing DIS methods' accuracy. Since stable diffusion generates masks by denoising from noise, modify the starting point using the latent features of an existing segmentation result patch instead, enabling the model to focus on refining details rather than rediscovering the overall structure.

Are existing DIS methods compatible with local patch inputs? To investigate this, we adopt the same window selection approach and patch replacement technique as default in LawDIS, and feed local patches[1] to each model to obtain masks. As illustrated in Tab. 6, except for ours, all earlier methods exhibit varying levels of performance decline. For example, the F_β^ω of UDUN [26] and BiRefNet [49] decreases by 3.8% and 3.9%, respectively. Additionally, Fig. 6 demonstrates the comparison of MVANet's segmentation on the full image (MVANet) versus local patches (MVANet‡), showing a notable performance drop for the latter setting. This indicates earlier methods relied on inputs of fixed resolutions do not accommodate variable inputs.

Methods	Fine-tuning VAE	Scheduler	Infer step	$F_\beta^\omega \uparrow$	$\mathcal{M} \downarrow$	FPS ↑
MVANet [45]	-	-	-	0.866	0.038	1.40
Ours-S	×	DDPM [11]	500	0.867	0.038	0.006
	×	DDIM [35]	10	0.835	0.044	0.55
	×	DDIM [35]	50	0.842	0.043	0.12
	×	TCD [48]	1	0.856	0.040	**3.09**
	✓	TCD [48]	1	**0.890**	**0.032**	3.07

Table 7. Efficiency analysis of LawDIS.

Efficiency analysis. We conduct an efficiency analysis using a single A100 GPU. First, we evaluate the results of denoising with the original DDPM scheduler for 500 steps, as shown in the 2nd row of Tab. 7. Then, we explore two acceleration strategies for the DDPM denoiser: the widely used DDIM denoiser [35] (3rd and 4th rows) and the one-step TCD denoiser [48] (5th row). Surprisingly, we find that the TCD denoiser largely preserves the segmentation performance of the DDPM denoiser while significantly improving efficiency, boosting FPS from 0.006 to 3.09. Furthermore, we fine-tune the VAE (6th row), which improves the model's ability to adapt to segmentation of high-resolution images without compromising efficiency. Finally, we compare the proposed LawDIS to the runner-up model MVANet [45], our model demonstrates significant advantages in both segmentation performance and inference speed, as shown in the 1st and 6th rows of Tab. 7.

5. Conclusion

We have presented a model named LawDIS that reformulates the DIS task within a generative diffusion framework by extending a single stable diffusion model into macro and micro modes, thus achieving controllable DIS in two aspects. In the macro mode, an LS strategy is proposed that enables the generation of segmentation results under the control of language prompts. In the micro mode, an WR strategy is designed to enable unlimited detail optimization in controllable windows at variable scales on high-resolution images. The two modes are integrated into a single network through a switcher, enabling collaborative enhancement during training and seamless mode-switching during inference. Extensive experiments on the DIS5K dataset demonstrate that our LawDIS significantly outperforms SOTA methods across all the metrics.

[1] Patches are resized to each model's required input resolution.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (NO.62376189 & NO.62476143). We express our sincere gratitude to Qi Ma (Nankai University) and Jingyi Liu (Keio University) for their constructive discussions.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 4, 2

[2] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023. 4, 2

[3] Bowen Cheng, Ross Girshick, Piotr Dollár, Alexander C Berg, and Alexander Kirillov. Boundary iou: Improving object-centric image segmentation evaluation. In *IEEE CVPR*, 2021. 7

[4] Ming-Ming Cheng and Deng-Ping Fan. Structure-measure: A new way to evaluate foreground maps. *IJCV*, 129(9): 2622–2638, 2021. 5

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 2

[6] Deng-Ping Fan, Ge-Peng Ji, Ming-Ming Cheng, and Ling Shao. Concealed object detection. *IEEE TPAMI*, 44(10): 6024–6042, 2021. 1

[7] Deng-Ping Fan, Ge-Peng Ji, Xuebin Qin, and Ming-Ming Cheng. Cognitive vision inspired object segmentation metric and loss function. *SSI*, 6(6):5, 2021. 5

[8] Deng-Ping Fan, Jing Zhang, Gang Xu, Ming-Ming Cheng, and Ling Shao. Salient objects in clutter. *IEEE TPAMI*, 45 (2):2344–2366, 2022. 1

[9] Deng-Ping Fan, Ge-Peng Ji, Peng Xu, Ming-Ming Cheng, Christos Sakaridis, and Luc Van Gool. Advances in deep concealed scene understanding. *VI*, 1(1):16, 2023. 1

[10] Stas Goferman, Lihi Zelnik-Manor, and Ayellet Tal. Context-aware saliency detection. *IEEE TPAMI*, 34(10): 1915–1926, 2011. 1

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 3, 5, 8

[12] Ge-Peng Ji, Guobao Xiao, Yu-Cheng Chou, Deng-Ping Fan, Kai Zhao, Geng Chen, and Luc Van Gool. Video polyp segmentation: A deep learning perspective. *MIR*, 19(6):531–549, 2022. 1

[13] Ge-Peng Ji, Deng-Ping Fan, Yu-Cheng Chou, Dengxin Dai, Alexander Liniger, and Luc Van Gool. Deep gradient learning for efficient camouflaged object detection. *MIR*, 20(1): 92–108, 2023. 1

[14] Ge-Peng Ji, Jingyi Liu, Peng Xu, Nick Barnes, Fahad Shahbaz Khan, Salman Khan, and Deng-Ping Fan. Frontiers in intelligent colonoscopy. *MIR*, 2025. 1

[15] 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 CVPR*, 2024. 5

[16] Taehun Kim, Kunhee Kim, Joonyeong Lee, Dongmin Cha, Jiho Lee, and Daijin Kim. Revisiting image pyramid structure for high resolution salient object detection. In *ACCV*, 2022. 1, 2, 6, 7, 8

[17] Hsin-Ying Lee, Hung-Yu Tseng, and Ming-Hsuan Yang. Exploiting diffusion prior for generalizable dense prediction. In *IEEE CVPR*, 2024. 3

[18] Ming Li, Taojiannan Yang, Huafeng Kuang, Jie Wu, Zhaoning Wang, Xuefeng Xiao, and Chen Chen. Controlnet++: Improving conditional controls with efficient consistency feedback. *ECCV*, 2024. 5

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 1

[20] Chang Liu, Xudong Jiang, and Henghui Ding. Primitivenet: decomposing the global constraints for referring segmentation. *VI*, 2(1):16, 2024. 1

[21] Chang Liu, Xiangtai Li, and Henghui Ding. Referring image editing: Object-level image editing via referring expressions. In *IEEE CVPR*, 2024. 1

[22] Feng Liu, Luan Tran, and Xiaoming Liu. Fully understanding generic objects: Modeling, segmentation, and reconstruction. In *IEEE CVPR*, 2021. 1

[23] Yun Liu, Yu-Huan Wu, Guolei Sun, Le Zhang, Ajad Chhatkuli, and Luc Van Gool. Vision transformers with hierarchical attention. *MIR*, 21(4):670–683, 2024. 1

[24] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *IEEE CVPR*, 2015. 2

[25] Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. How to evaluate foreground maps. In *IEEE CVPR*, 2014. 5

[26] 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 *ACM MM*, 2023. 1, 2, 6, 7, 8

[27] Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung. Saliency filters: Contrast based filtering for salient region detection. In *IEEE CVPR*, 2012. 5

[28] Haotian Qian, Yinda Chen, Shengtao Lou, Fahad Khan, Xiaogang Jin, and Deng-Ping Fan. Maskfactory: Towards high-quality synthetic data generation for dichotomous image segmentation. In *NeurIPS*, 2024. 2

[29] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In *IEEE CVPR*, 2019. 2, 5, 6

[30] Xuebin Qin, Zichen Zhang, Chenyang Huang, Masood Dehghan, Osmar R Zaiane, and Martin Jagersand. U2-net: Going deeper with nested u-structure for salient object detection. *PR*, 106:107404, 2020. 2, 5, 6

[31] Xuebin Qin, Deng-Ping Fan, Chenyang Huang, Cyril Diagne, Zichen Zhang, Adrià Cabeza Sant'Anna, Albert Suarez, Martin Jagersand, and Ling Shao. Boundary-aware segmentation network for mobile and web applications. *arXiv preprint arXiv:2101.04704*, 2021. 1

[32] Xuebin Qin, Hang Dai, Xiaobin Hu, Deng-Ping Fan, Ling Shao, and Luc Van Gool. Highly accurate dichotomous image segmentation. In *ECCV*, 2022. 1, 2, 5, 6, 7, 8

[33] 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 *ICML*, 2021. 4

[34] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE CVPR*, 2022. 2, 3, 4, 5, 7

[35] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 3, 4, 8

[36] Yuehao Song, Xinggang Wang, Jingfeng Yao, Wenyu Liu, Jinglin Zhang, and Xiangmin Xu. Vitgaze: gaze following with interaction features in vision transformers. *VI*, 2(1):1–15, 2024. 1

[37] Anne M Treisman and Garry Gelade. A feature-integration theory of attention. *COGPSY*, 12(1):97–136, 1980. 1

[38] Xiaoguang Tu, Zhi He, Yi Huang, Zhi-Hao Zhang, Ming Yang, and Jian Zhao. An overview of large ai models and their applications. *VI*, 2(1):1–22, 2024. 1

[39] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE TPAMI*, 43(10): 3349–3364, 2020. 2, 5, 6

[40] Mengyu Wang, Henghui Ding, Jun Hao Liew, Jiajun Liu, Yao Zhao, and Yunchao Wei. Segrefiner: Towards model-agnostic segmentation refinement with discrete diffusion process. *NeurIPS*, 2023. 2

[41] 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. 2, 3

[42] Chenxi Xie, Changqun Xia, Mingcan Ma, Zhirui Zhao, Xiaowu Chen, and Jia Li. Pyramid grafting network for one-stage high resolution saliency detection. In *IEEE CVPR*, 2022. 2, 6

[43] 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 *ICLR*, 2025. 2, 6, 7

[44] Haiyang Ying, Yixuan Yin, Jinzhi Zhang, Fan Wang, Tao Yu, Ruqi Huang, and Lu Fang. Omniseg3d: Omniversal 3d segmentation via hierarchical contrastive learning. In *IEEE CVPR*, 2024. 1

[45] Qian Yu, Xiaoqi Zhao, Youwei Pang, Lihe Zhang, and Huchuan Lu. Multi-view aggregation network for dichotomous image segmentation. In *IEEE CVPR*, 2024. 1, 2, 4, 6, 7, 8

[46] 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 *ICLR*, 2025. 2

[47] Yi Zeng, Pingping Zhang, Jianming Zhang, Zhe Lin, and Huchuan Lu. Towards high-resolution salient object detection. In *IEEE ICCV*, 2019. 1

[48] Jianbin Zheng, Minghui Hu, Zhongyi Fan, Chaoyue Wang, Changxing Ding, Dacheng Tao, and Tat-Jen Cham. Trajectory consistency distillation. *arXiv preprint arXiv:2402.19159*, 2024. 5, 8

[49] 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 AIR*, 3:9150038, 2024. 1, 2, 6, 7, 8

[50] Yan Zhou, Bo Dong, Yuanfeng Wu, Wentao Zhu, Geng Chen, and Yanning Zhang. Dichotomous image segmentation with frequency priors. In *IJCAI*, 2023. 1, 2, 4, 6

[51] Mingchen Zhuge, Deng-Ping Fan, Nian Liu, Dingwen Zhang, Dong Xu, and Ling Shao. Salient object detection via integrity learning. *IEEE TPAMI*, 45(3):3738–3752, 2022. 1