OMNIDIS: DICHOTOMOUS IMAGE SEGMENTATION WITH DETAIL-REFERENCE TOKENS

Anonymous authors
Paper under double-blind review

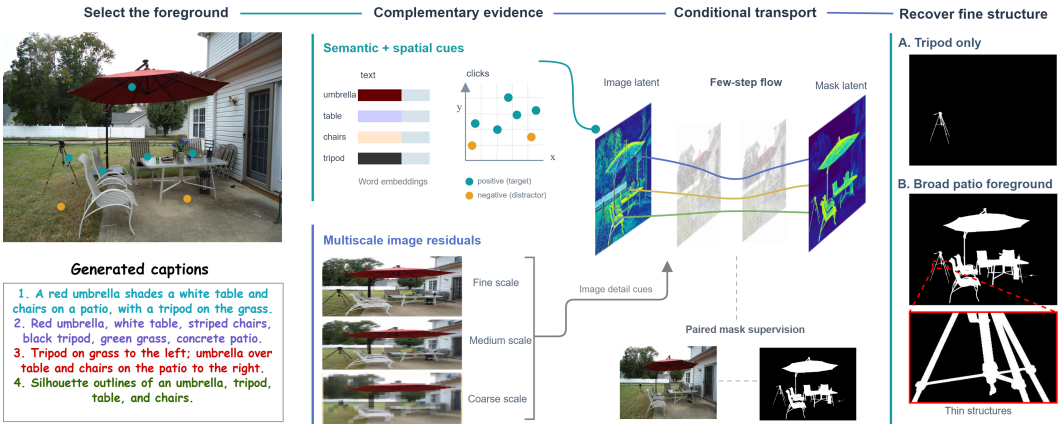

Figure 1: **Conceptual workflow of OmniDIS.** A description generated from the full image and labeled clicks specify the target through LPDP, while DRTI converts multiscale RGB residuals into visual reference tokens. Both streams condition image-to-mask flow before decoding.

ABSTRACT

High-fidelity dichotomous image segmentation requires both selecting the intended foreground and preserving its fine structure. Latent generative models offer strong visual priors, but compressed image representations can weaken boundary evidence, while a caption may match multiple objects. We present **OmniDIS**, which conditions image-to-mask flow on two complementary sources of evidence. Language–Point Dual-Prompting (LPDP) places semantic descriptions and labeled spatial cues in a shared context sequence, enabling text, point, and joint prompting with one model. Detail-Reference Token Injection (DRTI) converts multi-scale image residuals into learned reference tokens that condition the velocity field before decoding. This design lets target selection and detail cues interact during mask formation, without a separate refinement head. A frozen decoder transmits mask-supervision gradients to both modules, while low-rank adaptation updates the pretrained transformer. On DIS5K, OmniDIS-9B reaches $F_\beta^\omega = 0.949$ and $S_\alpha = 0.964$ on DIS-VD with four Euler steps. It surpasses all recent state-of-the-art methods, including FlowDIS and LawDIS, while OmniDIS-4B remains competitive with a substantially smaller generative backbone.

1 INTRODUCTION

Dichotomous image segmentation (DIS) estimates a detailed binary foreground mask from an image (Qin et al., 2022). Unlike coarse object localization, it must preserve narrow structures, holes, and intricate contours. These requirements matter in image editing, where an approximately correct object extent can still produce visible artifacts Goferman et al. (2011); Liu et al. (2021). Promptable DIS adds a second requirement: the mask should correspond to the user's intended foreground, which may differ from the most salient region. Accurate boundaries and accurate target selection must therefore be considered together.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Discriminative approaches improve detail through nested architectures, intermediate supervision, and specialized feature fusion (Qin et al., 2020; 2022; Pei et al., 2023; Zheng et al., 2024b; Yu et al., 2024; Kim et al., 2022). Generative approaches instead adapt pretrained image models to mask prediction (Yu et al., 2025; Yan et al., 2025; Sargsyan & Navasardyan, 2026). In particular, image-to-mask flow matching transports an image latent toward a mask latent, making a few-step generative predictor possible. However, a compressed image representation can attenuate local evidence, and a text description alone may not uniquely identify the requested target. Increasing backbone capacity does not by itself specify how these two kinds of missing evidence should enter the prediction process.

We ask: *how should a latent mask generator combine target-specific spatial evidence with image-derived detail cues?* A boundary detector alone cannot answer this question: textures and shadows also produce high-frequency responses. Likewise, a foreground click identifies a location but does not fully describe the object or its extent. We therefore supply these cues while the mask is being formed, so that the generative predictor can interpret local evidence in the context of the requested target.

We introduce **OmniDIS**, which adapts a pretrained FLUX.2 [klein] transformer Labs (2025) with two distinct conditioning interfaces. **Language–Point Dual-Prompting (LPDP)** expresses clicks as labeled, spatially indexed context tokens alongside language. The token content describes the cue, while its position identifies where that cue applies. **Detail-Reference Token Injection (DRTI)** summarizes multi-scale image residuals and maps them to the model's reference-token interface. Both influence the same velocity field, and decoded-mask supervision trains their interaction through a frozen VAE decoder. This allocation of semantic, spatial, and detail evidence lets a single latent predictor use target cues to interpret residual structure before decoding.

Our contributions are threefold:

- We formulate promptable DIS as image-to-mask flow in which target identity, labeled locations, and image-derived detail condition the latent prediction before decoding.
- We introduce LPDP to combine language and spatial point cues in the context stream, and DRTI to provide multiscale residual descriptors through a separate visual-reference stream.
- We evaluate 4B and 9B variants on DIS5K and use module, prompt, descriptor, reference-count, and sampling studies to identify where accuracy gains arise and what they cost.

2 Related Work

Dichotomous Image Segmentation. DIS5K provides detailed binary annotations for evaluating foreground segmentation (Qin et al., 2022; Xie et al., 2022; Zeng et al., 2019). U^2-Net (Qin et al., 2019; 2020) and IS-Net (Qin et al., 2022) use nested or intermediate representations; UDUN (Pei et al., 2023) separates trunk and structural information; BiRefNet (Zheng et al., 2024b) and MVANet (Yu et al., 2024) aggregate complementary spatial evidence. PDFNet Liu et al. (2026) instead augments RGB features with a pseudo-depth modality from monocular depth estimators Yang et al. (2024a;b); Bochkovskii et al. (2025), using a *depth integrity-prior* and cross-modal attention to sharpen boundaries without a generative backbone, following earlier RGB-D fusion designs Hu et al. (2024); Zhong et al. (2024). These methods motivate explicit treatment of fine structure, but their feature-processing designs differ from conditioning a pretrained latent velocity field.

Diffusion and Flow-Matching for Dense Prediction. Pretrained diffusion models can be repurposed for dense prediction (Xu et al., 2025; Ke et al., 2024). For DIS, DiffDIS exploits diffusion representations and auxiliary edge prediction (Yu et al., 2025), while LawDIS Yan et al. (2025) recasts DIS as macro (language-controlled) and micro (window-controlled) mask generation, distilled to a single sampling step via trajectory consistency distillation Zheng et al. (2024a). FlowDIS learns image-to-mask transport and improves text controllability using Position-Aware Instance Pairing (Sargsyan & Navasardyan, 2026). OmniDIS retains the image-to-mask flow formulation, adding direct click conditioning and residual-derived reference tokens. Its detail branch provides additional evidence to a learned predictor; it does not invert the VAE or guarantee recovery of every fine structure.

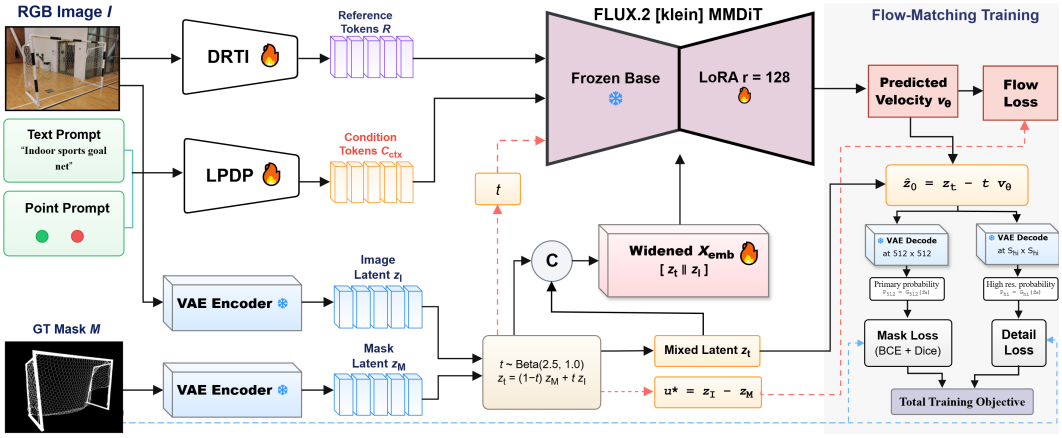

Figure 2: **OmniDIS training pipeline.** LPDP produces language–point context and DRTI produces image-derived reference tokens. The velocity predictor receives the mixed latent, the clean image latent, time, and both token streams. Flow supervision compares the predicted velocity with $z_I - z_M$. The one-step endpoint estimate is decoded separately at primary and higher resolutions for mask supervision. Decoding includes the probability readout; decoder weights are frozen while input gradients remain active.

Prompting and reference conditioning. SAM and SAM 2 support spatial prompting (Kirillov et al., 2023; Ravi et al., 2024), and SAMA explicitly addresses both segmentation and matting (Fan et al., 2026). DIS remains a binary-label task and is distinct from estimating a soft alpha matte. FLUX.2 [klein] already supports reference-conditioned generation and editing (Labs, 2025); rotary position embeddings (Su et al., 2021) also predate OmniDIS. We adapt these interfaces to a different role: language and clicks specify a segmentation target, while reference tokens convey descriptors derived from the same input image's residual structure. LPDP and DRTI remain separate input streams and interact through the mask predictor.

3 Proposed Method

3.1 Preliminaries: Image-to-Mask Flow

Let $I \in [0,1]^{H\times W\times 3}$ be an RGB image and $M \in \{0,1\}^{H\times W}$ its target mask. We write $\mathcal{E}$ for the frozen image-encoding interface, including input normalization and latent packing. The mask is replicated across three channels and normalized in the same way before encoding. At $H = W = 1024$, the packed latents $z_I = \mathcal{E}(I)$ and $z_M = \mathcal{E}(M)$ have shape $64 \times 64 \times 128$. This is the *packed* token grid: the released autoencoder first produces 32 channels at 128×128, then packs 2×2 neighborhoods into channels (Labs, 2025). Packing itself is reversible.

Following flow matching (Lipman et al., 2023) and its image-to-mask formulation (Sargsyan & Navasardyan, 2026), a training pair defines

$$z_t = (1-t)z_M + tz_I, \qquad u^\star = \frac{\mathrm{d}z_t}{\mathrm{d}t} = z_I - z_M, \tag{1}$$

where $t \sim \mathrm{Beta}(2.5, 1.0)$. The target $u^\star$ is constant for this *pair*; the learned conditional velocity field need not be constant. The timestep distribution emphasizes states near the image endpoint, from which inference starts. With context tokens c_{ctx} and detail references R, the flow objective is

$$\mathcal{L}_{\mathrm{fm}} = \mathbb{E}_{(I,M),t}\left[\frac{1}{n_z}\,\|v_\theta(z_t, z_I, t, c_{\mathrm{ctx}}, R) - u^\star\|_F^2\right], \tag{2}$$

where n_z normalizes the squared error over latent scalars.

3.1.1 Conditioning the Velocity Predictor

OmniDIS uses FLUX.2 [klein] with frozen base weights and trainable low-rank adapters (Labs, 2025; Hu et al., 2022). The main formulation uses the 9B variant with hidden width $d_h = 4096$. A

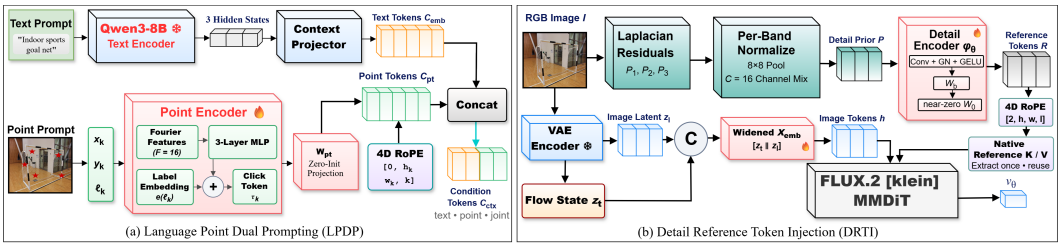

Figure 3: Distinct conditioning interfaces. Left: LPDP concatenates projected text and point content; spatial identifiers act on attention. Right: DRTI derives reference slots from RGB residuals and conditions the velocity predictor independently of LPDP.

widened input projection receives channel-concatenated flow and image latents:

$$h = x_{\text{emb}}([z_t \parallel z_I]), \qquad x^{(0)}_{\text{emb}} = [W_{\text{pre}} \;\; 0]. \tag{3}$$

This initialization preserves the original input projection response to z_t. Text is encoded once by a frozen Qwen3 encoder; the context projection maps the concatenated hidden states to $c_{\text{emb}} \in \mathbb{R}^{L_T \times d_h}$. LPDP adds point context after this projection, whereas DRTI enters the visual reference interface (Figure 2). Reference-token outputs are conditioning representations and are excluded from the mask-velocity target.

The one-step endpoint estimate used for training is

$$\hat{z}_0 = z_t - t\, v_\theta(z_t, z_I, t, c_{\text{ctx}}, R). \tag{4}$$

It equals z_M at the exact pairwise velocity; otherwise it supplies an approximate endpoint for decoded supervision (Appendix A).

3.2 Language–Point Dual-Prompting

Language–Point Dual-Prompting represents target identity and spatial constraints in one context sequence. Given a text prompt T and clicks $\mathcal{P} = \{(x_k, y_k, \ell_k)\}_{k=1}^{K}$, coordinates lie in $[0,1]^2$ and $\ell_k \in \{\text{fg}, \text{bg}\}$ specifies inclusion or exclusion. We sample two foreground and two background clicks uniformly without replacement from the geometrically augmented mask. Positive and negative labels indicate inclusion and exclusion, respectively (Appendix B.3).

Click content. For $F = 16$ frequency levels, let $\omega_f = 2^f$ and $(\tilde{x}_k, \tilde{y}_k) = s_c(x_k, y_k)$, with $s_c = 1$ (no additional coordinate scaling). The Fourier encoding is

$$\psi_k = [\sin(\omega_f \tilde{x}_k), \cos(\omega_f \tilde{x}_k), \sin(\omega_f \tilde{y}_k), \cos(\omega_f \tilde{y}_k)]_{f=0}^{F-1} \in \mathbb{R}^{4F}. \tag{5}$$

A GELU MLP with widths $64 \to 256 \to 256 \to d_h$ and a learned polarity embedding form

$$\tau_k = \text{MLP}(\psi_k) + e(\ell_k), \qquad c_{\text{pt}} = [(W_{\text{pt}}\tau_1)^\top; \ldots; (W_{\text{pt}}\tau_K)^\top], \quad c_{\text{ctx}} = [c_{\text{emb}}; c_{\text{pt}}]. \tag{6}$$

Here $\tau_k \in \mathbb{R}^{d_h}$, $W_{\text{pt}} \in \mathbb{R}^{d_h \times d_h}$, and semicolons concatenate rows in this equation. Coordinate features encode location, while the label embedding distinguishes inclusion from exclusion.

Spatial attention identifiers. Point tokens receive four-axis rotary identifiers $\rho_k = [0, h_k, w_k, k]$, with (h_k, w_k) the quantized click cell on the packed image-token grid. These identifiers are applied to attention queries and keys, following the backbone's rotary convention (Su et al., 2021; Labs, 2025). Together with token content, they expose *what*, *where*, and *include/exclude* cues during mask formation.

Prompt modes and initialization. With $K = 0$, no click tokens are appended. Point-only operation encodes an empty string as null-text context, and joint operation supplies both modalities through the same model interfaces. We initialize W_{pt} to zero, delaying the learned point-content contribution. Training drops text with probability 0.1 and clicks with probability 0.1. Appendix B details initialization and coordinate handling.

3.3 Detail-Reference Token Injection

Detail-Reference Token Injection (DRTI) conditions the velocity predictor on residual descriptors extracted from the input image. Because residuals also respond to texture and illumination, the predictor must interpret them together with the semantic and spatial target context.

Multi-scale residual descriptor. Starting from $I_1 = I$, let $I_{\ell+1} = \mathrm{Down}(I_\ell)$, where Down is $2\times$ average pooling. The residual at scale ℓ is

$$B_\ell = I_\ell - \mathrm{Up}(\mathrm{Down}(I_\ell)), \qquad \ell = 1, 2, 3, \tag{7}$$

with bilinear upsampling back to the size of I_ℓ. Each band is normalized and pooled to $g \times g$, $g = 8$. A channel-mixing operator A maps their channel-wise concatenation to

$$P = A([\mathrm{Pool}_g(\mathrm{Norm}(B_1)) \,\|\, \cdots \,\|\, \mathrm{Pool}_g(\mathrm{Norm}(B_3))]) \in \mathbb{R}^{16\times g\times g}. \tag{8}$$

The pooled P summarizes residual structure rather than preserving a full-resolution boundary map. It conditions learned mask synthesis without guaranteeing recovery of discarded spatial information.

Learned reference slots. A convolutional encoder with GroupNorm and GELU, followed by bottleneck and output projections, produces $N_R = 256$ tokens of width $d_z = 128$:

$$a = \mathrm{vec}(\mathrm{Conv}(P)), \qquad R = \mathrm{reshape}_{N_R\times d_z}(W_o W_b a). \tag{9}$$

The output projection has $N_R d_z = 32768$ outputs, mapping the pooled descriptor into learned slots rather than spatially upsampling it (Appendix C).

Reference-conditioned prediction. The tokens are projected through the visual reference input interface and assigned identifiers $[2, h_j^R, w_j^R, l_j^R]$. The first coordinate separates the reference namespace from the primary image tokens; it is not the flow time t. Reference information participates in the transformer's attention alongside image and context representations. The globally projected slots use positional identifiers without assuming one-to-one correspondence to image patches. The reference projection must accept width-128 inputs even though Eq. 3 widens the primary input to width 256.

Joint supervision. The output projection W_o is initialized at a small nonzero scale (10^{-3}). This limits initial content magnitude while permitting upstream gradients. DRTI learns through flow and decoded-mask supervision: R changes v_θ, which changes $\hat{z}_0$ and hence the decoded prediction. A frozen decoder carries these gradients without updating its weights. This conditions mask generation without a post-decoding refiner.

3.4 Objectives, Optimization, and Inference

Let $\mathcal{G}$ denote unpacking, frozen VAE decoding, conversion to one channel, and mapping to a valid foreground probability. We write $p_s = \mathcal{G}_s(\hat{z}_0)$ for the prediction evaluated on a spatial grid of size s and M_s for the aligned target. With ℓ_{seg} denoting BCE plus soft Dice loss,

$$\mathcal{L}_{\mathrm{mask}} = \ell_{\mathrm{seg}}(p_{512}, M_{512}), \qquad \mathcal{L}_{\mathrm{detail}} = \ell_{\mathrm{seg}}(p_{s_{\mathrm{hi}}}, M_{s_{\mathrm{hi}}}), \quad s_{\mathrm{hi}} > 512, \tag{10}$$

and the total objective is

$$\mathcal{L} = 0.1\mathcal{L}_{\mathrm{fm}} + \mathcal{L}_{\mathrm{mask}} + 0.1\mathcal{L}_{\mathrm{detail}}. \tag{11}$$

In the component ablation, the detail-loss coefficient is zero when DRTI is disabled; the factorial ablation in Appendix E varies these two factors independently. The higher-resolution term is ordinary segmentation supervision, not a discriminator or a frequency-specific loss. The decoded losses use separate decoder evaluations at their respective resolutions but share one velocity evaluation; Appendix C.3 defines the decoded losses.

We train the added input and prompt modules and the detail encoder, together with LoRA adapters on transformer attention and MLP projections. The base, text encoder, and VAE remain frozen. Images and masks share geometric augmentation; clicks are sampled afterward.

At inference, start from $z_1 = z_I$ and integrate backward on a decreasing grid $1 = t_0 > \cdots > t_J = 0$:

$$z_{t_{j+1}} = z_{t_j} + (t_{j+1} - t_j) v_\theta(z_{t_j}, z_I, t_j, c_{\mathrm{ctx}}, R), \quad j = 0, \ldots, J-1. \tag{12}$$

The final state is decoded once to obtain the mask probability. We report $J = 4$ by default, with one- and two-step results for comparison. We use the native backbone sampling schedule; the Beta distribution in Eq. 1 governs training-time sampling only (Appendix A).

Table 1: Results on **dichotomous image segmentation** (DIS5K). ↑: higher is better; ↓: lower is better. **Bold**: best; Underline: second best. †: Language/text guided model.

Method	Venue	DIS-VD (470)					DIS-TE1 (500)					DIS-TE2 (500)				
		$S_\alpha\uparrow$	$F_\beta^{mx}\uparrow$	$F_\beta^{w}\uparrow$	$E_\phi^{mn}\uparrow$	$\mathcal{M}\downarrow$	$S_\alpha\uparrow$	$F_\beta^{mx}\uparrow$	$F_\beta^{w}\uparrow$	$E_\phi^{mn}\uparrow$	$\mathcal{M}\downarrow$	$S_\alpha\uparrow$	$F_\beta^{mx}\uparrow$	$F_\beta^{w}\uparrow$	$E_\phi^{mn}\uparrow$	$\mathcal{M}\downarrow$
U^2Net	PR_{20}	0.785	0.753	0.656	0.809	0.089	0.762	0.701	0.601	0.783	0.085	0.798	0.768	0.676	0.825	0.083
IS-Net	$CVPR_{22}$	0.813	0.791	0.717	0.856	0.074	0.787	0.740	0.662	0.820	0.074	0.823	0.799	0.728	0.858	0.070
UDUN	MM_{23}	0.838	0.823	0.763	0.892	0.059	0.817	0.784	0.720	0.864	0.059	0.843	0.829	0.768	0.886	0.058
BiRefNet	$CAAI_{24}$	0.898	0.891	0.854	0.931	0.038	0.885	0.860	0.819	0.911	0.037	0.900	0.893	0.857	0.930	0.036
PDFNet	$CVPR_{26}$	0.916	0.913	0.875	0.945	0.027	0.899	0.888	0.848	0.928	0.030	0.924	0.919	0.886	0.948	0.027
SAMA	$AAAI_{26}$	0.930	0.942	0.885	0.962	0.021	0.947	0.940	0.911	0.977	0.012	0.934	0.932	0.904	0.962	0.019
LawDIS-R†	$ICCV_{25}$	0.925	0.929	0.905	0.959	0.023	0.917	0.919	0.890	0.948	0.024	0.932	0.936	0.914	0.961	0.022
FlowDIS†	$CVPR_{26}$	0.953	0.958	0.938	0.974	0.014	0.953	0.961	0.942	0.976	0.012	0.958	0.965	0.947	0.978	0.012
OmniDIS-4B	–	0.951	0.956	0.935	0.971	0.015	0.954	0.959	0.939	0.970	0.014	0.955	0.964	0.948	0.975	0.013
OmniDIS-9B	–	**0.964**	**0.970**	**0.949**	**0.987**	**0.012**	**0.964**	**0.970**	**0.949**	**0.982**	**0.010**	**0.965**	**0.973**	**0.959**	**0.982**	**0.011**

Method	Venue	DIS-TE3 (500)					DIS-TE4 (500)					DIS-TE1–4 (Avg.) (2,000)				
		$S_\alpha\uparrow$	$F_\beta^{mx}\uparrow$	$F_\beta^{w}\uparrow$	$E_\phi^{mn}\uparrow$	$\mathcal{M}\downarrow$	$S_\alpha\uparrow$	$F_\beta^{mx}\uparrow$	$F_\beta^{w}\uparrow$	$E_\phi^{mn}\uparrow$	$\mathcal{M}\downarrow$	$S_\alpha\uparrow$	$F_\beta^{mx}\uparrow$	$F_\beta^{w}\uparrow$	$E_\phi^{mn}\uparrow$	$\mathcal{M}\downarrow$
U^2Net	PR_{20}	0.823	0.813	0.721	0.856	0.073	0.814	0.800	0.707	0.837	0.085	0.799	0.771	0.676	0.825	0.082
IS-Net	$CVPR_{22}$	0.836	0.830	0.758	0.883	0.064	0.830	0.827	0.753	0.870	0.072	0.819	0.799	0.726	0.858	0.070
UDUN	MM_{23}	0.865	0.865	0.809	0.917	0.050	0.849	0.846	0.792	0.901	0.059	0.844	0.831	0.772	0.892	0.057
BiRefNet	$CAAI_{24}$	0.919	0.925	0.893	0.955	0.028	0.900	0.904	0.864	0.939	0.039	0.901	0.891	0.858	0.934	0.035
PDFNet	$CVPR_{26}$	0.928	0.929	0.901	0.957	0.027	0.909	0.910	0.869	0.941	0.037	0.916	0.915	0.876	0.944	0.030
SAMA	$AAAI_{26}$	0.924	0.920	0.889	0.949	0.032	0.897	0.917	0.857	0.937	0.041	0.925	0.926	0.897	0.956	0.026
LawDIS-R†	$ICCV_{25}$	0.932	0.929	0.919	0.964	0.022	0.922	0.922	0.910	0.964	0.026	0.926	0.932	0.905	0.959	0.023
FlowDIS†	$CVPR_{26}$	0.954	0.963	0.942	0.975	0.015	0.939	0.946	0.919	0.964	0.024	0.951	0.959	0.938	0.973	0.016
OmniDIS-4B	–	0.952	0.961	0.942	0.973	0.016	0.941	0.949	0.915	0.961	0.026	0.950	0.958	0.936	0.970	0.017
OmniDIS-9B	–	**0.963**	**0.970**	**0.951**	**0.988**	**0.012**	**0.952**	**0.958**	**0.924**	**0.969**	**0.020**	**0.961**	**0.968**	**0.946**	**0.980**	**0.013**

4 EXPERIMENTS

4.1 SETUP AND IMPLEMENTATION DETAILS

DIS5K contains 5470 image–mask pairs across 225 categories, with 3000 training images, 470 validation images, and four test subsets of 500 images each (Qin et al., 2022). We train on DIS-TR and evaluate on DIS-VD and the four DIS-TE subsets. For each full image, a Qwen3 VLM generates a one-to-three-sentence target description offline, without target crops or mask-assisted input. Each image is associated with four paraphrased prompts, from which one is sampled per training step; frozen text-encoder embeddings are precomputed offline. Caption generation is separate from LPDP text encoding (Appendix B.1).

We evaluate FLUX.2 [klein] 4B and 9B with their native frozen Qwen3-4B and Qwen3-8B text encoders, respectively. Training uses 4 NVIDIA H100 GPUs (94 GB each), a total batch size of 32, and 10,000 updates, with a training time of approximately 5 hour. Inputs are processed at 1024×1024 with random flips and 90° rotations. We use AdamW with an initial learning rate of 5×10^{-5}, halved at updates 512, 2048, 4096, and 8192. Training timesteps follow $\mathrm{Beta}(2.5, 1.0)$. Rank-128 LoRA adapts the attention and MLP projections; the added input, point-conditioning, and detail-reference components are trained jointly, while the VAE and text encoder remain frozen. The flow, mask, and auxiliary detail losses have weights 0.1, 1.0, and 0.1, respectively. Inference uses 4 Euler steps with the backbone's native sampling schedule. Appendix D provides the adaptation configuration and model-specific interfaces.

We report structure measure S_α, maximum F-measure F_β^{mx}, weighted F-measure F_β^{ω}, mean enhanced-alignment measure E_ϕ^{mn}, and mean absolute error $\mathcal{M}$. Higher is better except for $\mathcal{M}$. Evaluation simulates four fixed ground-truth-mask clicks, sampled once with seed 42 and cached across click-enabled comparisons. The prompt-mode study partitions DIS-VD into 370 single-object and 100 multi-object images.

4.2 MAIN RESULTS

Table 1 compares OmniDIS with discriminative methods, including BiRefNet (Zheng et al., 2024b) and PDFNet (Liu et al., 2026), and generative approaches such as LawDIS (Yan et al., 2025) and FlowDIS (Sargsyan & Navasardyan, 2026). Gains below are absolute differences on a 100-point scale. On DIS-VD, OmniDIS-9B achieves $F_\beta^{\omega} = 0.949$ and $S_\alpha = 0.964$, exceeding PDFNet and BiRefNet by 7.4 and 9.5 points in F_β^{ω}, respectively. Across the four test subsets, it achieves a macro-average $F_\beta^{\omega} = 0.946$, outperforming PDFNet by 7.0 points and FlowDIS by 0.8 points. Its macro-

Table 2: Component results. "Baseline" is the DIS adaptation without LPDP or DRTI.

Ablation settings	$F_\beta^w \uparrow$	$F_\beta^{mx} \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$E_\phi^{mn} \uparrow$
Baseline Labs (2025)	0.862	0.911	0.017	0.896	0.931
w/ DRTI	0.921	0.948	0.013	0.937	0.964
w/ LPDP	0.880	0.923	0.015	0.910	0.945
OmniDIS-9B	**0.949**	**0.970**	**0.012**	**0.964**	**0.987**

Table 3: LoRA rank on OmniDIS-9B. T.P = Trainable Parameters.

r	$F_\beta^\omega \uparrow$	$F_\beta^{mx} \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$E_\phi^{mn} \uparrow$	T.P (M)
32	0.933	0.957	0.014	0.951	0.974	49.65
64	0.942	0.963	0.013	0.958	0.981	81.06
128	**0.949**	**0.970**	**0.012**	**0.964**	**0.987**	143.88

Table 4: Ablation of DRTI vs. other refinement.

Strategy	$F_\beta^\omega \uparrow$	$F_\beta^{mx} \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$
(a) No refinement	0.862	0.911	0.017	0.896
(b) VAE enc. feats	0.845	0.892	0.021	0.881
(c) DRTI feats.	0.881	0.933	0.015	0.911
(d) **DRTI (in-latent)**	**0.921**	**0.948**	**0.013**	**0.937**

Table 5: Ablation on LPDP prompt modes.

Mode	DIS-VD-Single (370)				DIS-VD-Multi (100)			
	$F_\beta^\omega \uparrow$	$F_\beta^{mx} \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$F_\beta^\omega \uparrow$	$F_\beta^{mx} \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$
Text-only	0.949	0.963	0.012	0.958	0.913	0.939	0.015	0.933
Point-only	0.947	0.960	0.013	0.959	0.924	0.948	0.014	0.941
Text+Point	**0.952**	**0.973**	**0.012**	**0.966**	**0.938**	**0.960**	**0.012**	**0.956**

average $S_\alpha = 0.961$ also exceeds FlowDIS by 1.0 point. These results support the effectiveness of conditioning mask generation on both target identity and image-derived detail; the component analysis below examines their individual contributions.

Consistency across subsets. OmniDIS-9B improves F_β^ω over FlowDIS on all four test subsets, with gains ranging from 0.5 to 1.2 points. On the challenging DIS-TE4 subset, it improves S_α by 1.3 points and reduces $\mathcal{M}$ from 0.024 to 0.020. The gains span all subsets and are consistent with improved structural prediction.

Model scale. OmniDIS-9B improves F_β^ω over OmniDIS-4B by 0.9–1.1 points across the four test subsets. FlowDIS uses the 12B FLUX.1-schnell transformer (Sargsyan & Navasardyan, 2026). OmniDIS-4B reaches a test macro-average $F_\beta^\omega = 0.936$, within 0.2 points of FlowDIS at 0.938, using roughly one-third of the generative-backbone parameters. This comparison excludes text encoders and does not equate pretraining or conditioning. The 9B variant favors accuracy; the 4B variant offers a smaller-backbone alternative.

Figure 4 compares predictions on thin structures and complex foregrounds, while Figure 5 illustrates language-conditioned target selection.

4.3 Ablation Studies

Table 2 compares the baseline without either module, DRTI alone, LPDP alone, and the full OmniDIS-9B model. Relative to the baseline, DRTI alone improves F_β^ω by 5.9 points, while LPDP alone improves it by 1.8 points. Combining both modules raises F_β^ω from 0.862 to 0.949, an improvement of 8.7 points. The full model also exceeds DRTI alone by 2.8 points and LPDP alone by 6.9 points, indicating that each module contributes beyond the other. These results are consistent with complementary roles: LPDP conditions prediction on the requested target, while DRTI supplies image-derived detail cues.

DRTI: contribution and injection. Using DRTI alone raises F_β^ω from 0.862 to 0.921 and S_α from 0.896 to 0.937, corresponding to gains of 5.9 and 4.1 points over the baseline. When combined with LPDP, DRTI provides a further 6.9-point improvement in F_β^ω over the LPDP-only configuration. Appendix E separates reference conditioning from auxiliary supervision. With LPDP fixed, references without the detail loss improve F_β^ω by 5.9 points; adding the loss then contributes 1.0 point. The injection comparison in Table 4 evaluates DRTI without LPDP: it improves over the DRTI-feature refiner by 4.0 points in F_β^ω (0.881 vs. 0.921) and 2.6 points in S_α (0.911 vs. 0.937), while reducing $\mathcal{M}$ from 0.015 to 0.013, and leads all reported metrics in Table 4. This supports routing residual-derived detail into the latent predictor rather than applying it only as feature-level refinement.

LPDP: language and spatial cues. LPDP alone improves baseline F_β^ω by 1.8 points, and adding it to the DRTI-only configuration yields a further 2.8-point gain (Table 2). Table 5 compares lan-

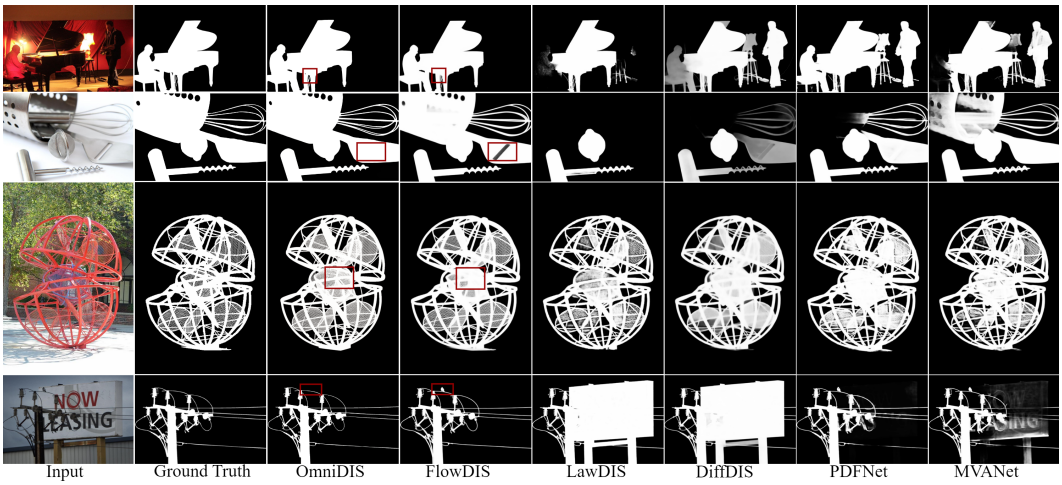

Figure 4: **Qualitative DIS5K comparisons.** Each row shows the input, reference mask, and predictions from OmniDIS and the listed methods. Red boxes mark selected fine structures and regions where the outputs differ.

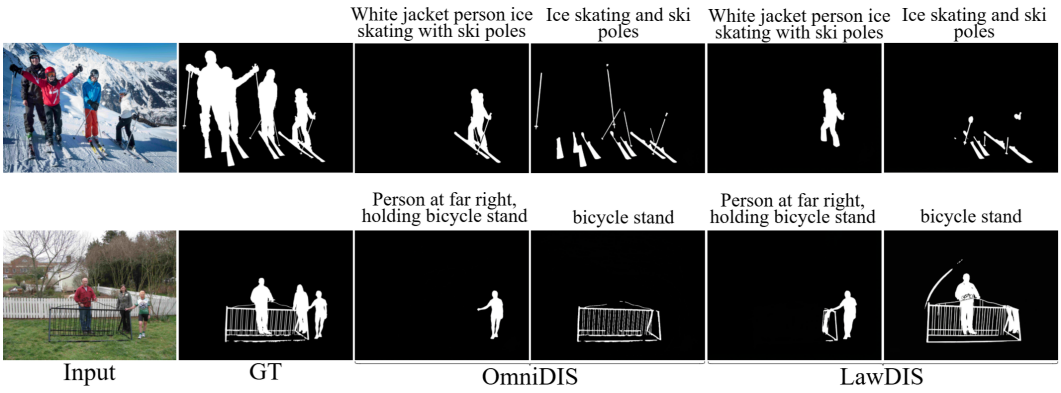

Figure 5: **Language-conditioned target selection.** Different descriptions of the same scene produce different OmniDIS and LawDIS masks. The displayed GT is the dataset reference mask; separate ground-truth masks are not shown for every description.

guage and clicks within the full model. Joint prompting improves F_β^ω over the stronger individual modality by 0.3 points on the single-object group and 1.4 points on the multi-object group. On multi-object images, points alone also exceed text alone by 1.1 points. The larger joint-prompt gain in multi-object scenes is consistent with spatial cues helping distinguish the requested target when a description is ambiguous. Joint prompting achieves the highest F_β^ω, F_β^{mx}, and S_α in both groups; it lowers $\mathcal{M}$ on multi-object images and ties text-only on single-object images. The encoding ablation in Appendix E shows that removing spatial rotary identifiers or continuous coordinate features reduces F_β^ω by 1.0 or 1.5 points, respectively, supporting the use of both representations.

Descriptor resolution and reference count. Table 6 and 7 examine two ways of controlling the DRTI interface. With 256 reference slots, increasing the pooling grid from 4×4 to 8×8 raises F_β^w from 0.940 to 0.949; 16×16 reaches 0.950 but slightly lowers F_β^{mx}, so we retain the compact 8×8 grid. At that grid size, increasing the number of slots from 64 to 128 and then 256 raises F_β^w by 0.7 and 0.5 points, respectively. The gain also changes the output-projection size, so it should be interpreted as a reference-capacity comparison rather than an isolated effect of token count.

Table 6: DRTI pooling grids with residual bands.

Proposed grid	$F_\beta^\omega \uparrow$	$F_\beta^{mx} \uparrow$	$S_\alpha \uparrow$	$\mathcal{M} \downarrow$
4×4	0.940	0.963	0.957	0.014
8×8	0.949	0.970	0.964	0.012
16×16	0.950	0.969	0.964	0.012

Table 7: DRTI reference count.

Reference slots	$F_\beta^\omega \uparrow$	$F_\beta^{mx} \uparrow$	$S_\alpha \uparrow$	$\mathcal{M} \downarrow$
64	0.937	0.960	0.953	0.014
128	0.944	0.966	0.960	0.013
256	0.949	0.970	0.964	0.012

Parameter Efficient Training. Table 3 compares adaptation ranks on the frozen 9B backbone. Rank 128 improves F_β^ω by 1.6 points over rank 32 and gives the highest tested accuracy. Its 143.88M trainable parameters include all adapters and added modules, approximately 1.60% of the nominal backbone size.

Table 8: Efficiency analysis of OmniDIS on DIS-TE4.

Method	Scheduler	Infer Step	FPS ↑	$F_\beta^\omega \uparrow$	$\mathcal{M} \downarrow$
LawDIS	TCD	1	3.09	0.890	0.032
LawDIS	DDIM	50	0.12	0.842	0.043
OmniDIS-4B	Euler	1	**6.93**	0.908	0.029
OmniDIS-9B	Euler	1	4.61	0.916	0.022
		2	3.71	0.919	0.021
		4	2.60	**0.924**	**0.020**

Sampling Cost and Limitations. Increasing OmniDIS-9B from one to four Euler steps improves F_β^ω by 0.8 points on DIS-TE4 and reduces $\mathcal{M}$ from 0.022 to 0.020, while throughput decreases from 4.61 to 2.60 FPS (Table 8). The one-step 4B variant provides a faster operating point of 6.93 FPS at $F_\beta^\omega = 0.908$. LawDIS uses TCD (Zheng et al., 2024a) and DDIM (Song et al., 2021) under different setups, precluding a matched-hardware speedup claim. Evaluation remains confined to DIS5K and fixed simulated clicks, leaving cross-dataset generalization and robustness to inaccurate or adaptive clicks untested. Although the resolution and point-encoding ablations examine two design choices, compact descriptors may still omit local structure, and strong background residuals or conflicting prompts can misdirect target selection. These limitations motivate matched component controls and evaluation under perturbed prompts and unfamiliar image distributions. Graphic-content failures are discussed in Appendix F.

5 Conclusion

OmniDIS addresses two limitations of latent mask generation: an image latent does not specify which foreground is requested, and compression can weaken evidence for fine structure. LPDP supplies target descriptions and labeled locations as context tokens; DRTI supplies compact residual-derived references through the backbone's native visual-reference path. This separation matters because a high-frequency pattern may be a relevant boundary for one target and background texture for another. Conditioning the velocity field lets the model interpret both cues while constructing the mask, before frozen VAE decoding. The component controls indicate that the benefit of reference tokens is not simply an effect of adding another decoded loss. The prompt-mode study further shows how spatial cues can help when descriptions admit several interpretations. These findings must be read alongside the information supplied by the evaluation protocol: descriptions are generated offline from full images, and evaluation clicks are sampled from ground-truth masks rather than provided by users. Residual pooling also compresses local evidence; disconnected strokes, narrow gaps, and textured backgrounds can therefore remain difficult even when the correct target is selected. The present study establishes this conditioning interface on DIS5K, but does not establish robustness to imprecise interaction or transfer to unfamiliar image distributions.

Future work will evaluate real and perturbed click sequences, determine when further interaction is useful, and develop spatially structured references that retain thin components without indiscriminately increasing the token budget. We will test these changes across unfamiliar image distributions and at higher mask resolutions.

AI USE STATEMENT

In this work, we used generative AI tools for coding tasks, specifically Claude Code and Codex coding agents. We have also used generative AI such as Qwen3 VL for image caption generation (data generations) for training our model. We have not used generative AI tools for other tasks requiring disclosure, such as research ideation, experimental design, or generation of scientific claims; the remaining required-disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for language polishing and editing of the manuscript, which is a recommended-disclosure task. We have reviewed all AI-assisted work. In particular, AI-generated code was inspected, adapted, and tested for correctness; AI-polished text was checked for accuracy, originality, and consistency with our intended meaning.

ETHICS STATEMENT

This work does not involve human subjects, personally identifiable information, or sensitive data. The experiments are conducted on publicly available benchmark datasets and do not involve applications that we believe raise ethical concerns. We do not foresee any negative societal impacts specific to this work. To the best of our knowledge, this manuscript does not cover any topics or experiments that require ethical review under the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, this manuscript provides detailed implementation details, model configurations, hyperparameter settings, evaluation metrics, and experimental protocols. We also specify the computational resources and random seeds used where applicable. The code will be publicly released upon acceptance.

REFERENCES

Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R. Richter, and Vladlen Koltun. Depth Pro: Sharp monocular metric depth in less than a second. In *International Conference on Learning Representations*, 2025.

Zezhong Fan, Xiaohan Li, Topojoy Biswas, Kaushiki Nag, and Kannan Achan. Segment and matte anything in a unified model. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2026.

Stas Goferman, Lihi Zelnik-Manor, and Ayellet Tal. Context-aware saliency detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(10):1915–1926, 2011.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2106.09685.

Xihang Hu, Fuming Sun, Jing Sun, Fasheng Wang, and Haojie Li. Cross-modal fusion and progressive decoding network for rgb-d salient object detection. *International Journal of Computer Vision*, pp. 1–19, 2024.

Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

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 (ACCV)*, pp. 108–124, 2022.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4015–4026, 2023.

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2, 2025.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Feng Liu, Luan Tran, and Xiaoming Liu. Fully understanding generic objects: Modeling, segmentation, and reconstruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Xianjie Liu, Keren Fu, and Qijun Zhao. High-precision dichotomous image segmentation via depth integrity-prior and fine-grained patch strategy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

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 (ACM MM)*, pp. 2139–2147, 2023.

Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

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. *Pattern Recognition*, 106:107404, 2020.

Xuebin Qin, Hang Dai, Xiaobin Hu, Deng-Ping Fan, Ling Shao, and Luc Van Gool. Highly accurate dichotomous image segmentation. In *European Conference on Computer Vision (ECCV)*, pp. 38–56, 2022.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

Andranik Sargsyan and Shant Navasardyan. Flowdis: Language-guided dichotomous image segmentation with flow matching. *arXiv preprint arXiv:2605.05077*, 2026.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021. URL https://arxiv.org/abs/2104.09864.

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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11717–11726, 2022.

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 *International Conference on Learning Representations (ICLR)*, 2025.

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 (ICCV)*, 2025.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024a.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024b.

Qian Yu, Xiaoqi Zhao, Youwei Pang, Lihe Zhang, and Huchuan Lu. Multi-view aggregation network for dichotomous image segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3921–3930, 2024.

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

Yi Zeng, Pingping Zhang, Jianming Zhang, Zhe Lin, and Huchuan Lu. Towards high-resolution salient object detection. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

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

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, 2024b.

Mingyu Zhong, Jing Sun, Peng Ren, Fasheng Wang, and Fuming Sun. Magnet: Multi-scale awareness and global fusion network for rgb-d salient object detection. *Knowledge-Based Systems*, pp. 112126, 2024.

A Flow Direction and Decoded Supervision

Paired target versus learned field. The path in Eq. 1 has derivative $u^\star = z_I - z_M$, constant for a particular image–mask pair. The shared network v_θ is trained against these pairwise targets and need not be constant along an inferred trajectory. Writing $e_t = v_\theta - u^\star$, the endpoint estimate in Eq. 4 satisfies

$$\hat{z}_0 - z_M = -te_t, \qquad \|\hat{z}_0 - z_M\|_F^2 = t^2\|e_t\|_F^2. \tag{13}$$

The estimate is exact at the target velocity, but one-step training supervision is not an exact inference solver. Near $t = 1$, the predictor must infer the mask from a state dominated by the image; training with $t \sim \mathrm{Beta}(2.5, 1)$ emphasizes that regime. At $t = 0$, endpoint supervision supplies no velocity gradient because the state already equals z_M.

Inference direction and schedule. Inference starts from $z_1 = z_I$, without added Gaussian noise, and uses explicit Euler integration on descending timesteps. Thus $t_{j+1} - t_j < 0$ in Eq. 12. At the exact pairwise velocity, a single step from 1 to 0 gives $z_I - (z_I - z_M) = z_M$. For learned velocities, additional steps reevaluate the field at intermediate states. The experiments use the native backbone sampling schedule, with four steps by default. The Beta distribution defines training-time sampling and is not an inverse-CDF inference grid.

Gradients through frozen decoding. Let η denote the DRTI parameters and $J_{f,x}$ the Jacobian of f with respect to x. For $p_s = \mathcal{G}_s(\hat{z}_0)$ and a decoded loss $\ell(p_s, M_s)$,

$$\nabla_\eta \ell = -t\, J_{R,\eta}^\top J_{v,R}^\top J_{\mathcal{G}_s,\hat{z}_0}^\top \nabla_{p_s}\ell. \tag{14}$$

Freezing the decoder parameters retains its input Jacobian, allowing supervision to reach DRTI, LPDP, and the trainable adapters. Thresholding is reserved for visualization or threshold-dependent evaluation; it is not part of this differentiable loss path.

B Language–Point Conditioning

B.1 Image Descriptions and Text Encoding

A Qwen3 vision–language model processes each full DIS5K image offline to generate a description of its main target. Caption generation uses neither target-cropped images nor ground-truth-mask-assisted views. Descriptions contain one to three sentences and express object identity, appearance, or relative position. Each image is associated with four paraphrased descriptions; one is selected per training step. The descriptions provide image-specific appearance and spatial context beyond a category label.

Caption generation and segmentation-time text encoding are distinct operations. The captioning VLM receives the full RGB image. LPDP's frozen Qwen3 text encoder receives only the resulting text; it does not receive the RGB image. The image reaches the segmentation predictor through the VAE and DRTI branches. Text embeddings are precomputed offline. Consequently, the language-guided system uses additional pretrained semantic processing compared with an RGB-only baseline, even when the description is derived from the image. Cross-method scores are therefore interpreted under each method's conditioning protocol, without an identical-input superiority claim.

B.2 Point Content and Spatial Identifiers

A click supplies both a location and an inclusion/exclusion label. LPDP represents these through continuous Fourier features and a learned polarity embedding, then supplies a spatial identifier for rotary attention. The two encodings play different roles: content distinguishes coordinates and polarity, while identifiers determine the relative positional transformation of attention queries and keys.

We use $s_c = 1$: the normalized coordinates enter Eq. 5 directly, with no additional scale or 2π factor. Sixteen frequencies $\omega_f = 2^f$ produce 64 features. The 9B PointEncoder uses widths $64 \to 256 \to 256 \to 4096$ with GELU nonlinearities, a learned 4096-dimensional polarity embedding,

and the point projection W_{pt}. Quantization affects only the rotary identifiers; the content features retain continuous coordinates.

For the 9B model, the concatenated text features have width 12288 and the pretrained context projection maps them to $d_h = 4096$. The point projection outputs the same hidden width, so its tokens are appended after the text context projection. The corresponding released 4B widths are 7680 and 3072. Point tokens must match the selected transformer's hidden width rather than the unprojected text-feature width.

B.3 Point Sampling and Geometric Consistency

For a target mask M, define the eligible pixel sets

$$\Omega_{\mathrm{fg}} = \{(u, v) : M[v, u] = 1\}, \qquad \Omega_{\mathrm{bg}} = \{(u, v) : M[v, u] = 0\}. \tag{15}$$

We draw $K_{\mathrm{fg}} = 2$ foreground and $K_{\mathrm{bg}} = 2$ background pixels uniformly without replacement within each set, giving $K = 4$ clicks. No erosion margin or background exclusion band is applied. Positive labels indicate inclusion in the requested mask, so an unselected object can provide a negative click even when it shares the target's category.

Training first applies random flips and 90° rotations jointly to the image and mask. Clicks are then sampled from the transformed mask, avoiding a separate coordinate transform. A sampled pixel (u, v) is represented by its normalized center,

$$(x_k, y_k) = \left(\frac{u + 0.5}{W}, \frac{v + 0.5}{H} \right), \qquad h_k = \min(63, \lfloor 64 y_k \rfloor), \quad w_k = \min(63, \lfloor 64 x_k \rfloor). \tag{16}$$

The grid is the packed 64×64 image-token grid. Foreground clicks precede background clicks; within each polarity, the order is random. Sequence concatenation preserves this order, and k indexes the token row and fourth rotary coordinate consistently.

Simulated interaction at evaluation. We sample two foreground and two background clicks once from each ground-truth mask with seed 42, save their coordinates, and reuse them across evaluated methods that accept clicks. Clicks are fixed rather than accumulated through iterative correction. This protocol measures segmentation under simulated interaction; it supplies target-specific spatial information beyond the RGB image and generated description. Text-only methods receive no clicks. Table 5 partitions the 470 validation images into 370 single-object and 100 multi-object cases.

Prompt availability. For point-only prediction, the frozen Qwen3 encoder processes an empty string to obtain null-text context tokens. During training, text is dropped with probability 0.1 and clicks with probability 0.1. Dropped text uses null-text context; dropped clicks give an empty point sequence. Uniform sampling does not ensure coverage of every thin structure or disconnected foreground component.

B.4 Attention and Initialization

For query Q_i, key K_j, and rotary transformation $\mathcal{R}$, the positional contribution enters through

$$\big[\mathcal{R}(\rho_i) Q_i\big]^\top \big[\mathcal{R}(\rho_j) K_j\big]. \tag{17}$$

Content and position jointly determine the logit; spatial proximity alone does not guarantee higher attention. Click-order identifiers also mean that permutation invariance of the click sequence is not assumed. With no clicks, the appended point sequence is empty; joint prompting includes both text and labeled points.

Zero initialization of W_{pt} limits initial point content but does not mask out tokens. In the simplified case of K added zero-key, zero-value tokens, an original attention output S/Z becomes $S/(Z + K)$. Thus the initialization does not exactly preserve pretrained attention. Initially, upstream point layers receive zero gradient through a zero projection, while W_{pt} can receive a gradient and subsequently enable their learning.

C Detail References and Decoder Losses

C.1 Descriptor and Token Dimensions

Three RGB residual bands supply nine channels. After normalization and pooling, channel mixing maps $9 \times g \times g$ values to $16 \times g \times g$, with $g = 8$ by default. Let the flattened convolutional features contain q scalars and the bottleneck width be b. Then

$$W_b \in \mathbb{R}^{b \times q}, \qquad W_o \in \mathbb{R}^{N_R d_z \times b}, \qquad R \in \mathbb{R}^{N_R \times d_z}, \tag{18}$$

where $N_R = 256$ and $d_z = 128$. The learned output projection reshapes into reference slots; it is not a spatial upsampler from an 8×8 grid to 256 preserved patches. Because the pooled descriptor has 1024 scalars, its local differentiable map to R has Jacobian rank at most 1024. This observation applies only to the descriptor branch, not to the full predictor, which also receives image latents and target context.

The widened primary input has 256 values per packed token, whereas the reference input remains width 128. Their input projections therefore have distinct dimensional roles. Reference identifiers $[2, h_j^R, w_j^R, l_j^R]$ separate the reference namespace from the primary image tokens. The first coordinate is not flow time. A slot's assigned spatial identifier is a positional convention, not a guarantee that it encodes one corresponding image patch. The reference slots enter the backbone's native visual-reference key/value pathway. Because R is fixed for an input image, its reference-side keys and values are computed once and reused across Euler evaluations, while the evolving flow state and velocity prediction are recomputed at every step. This reuse is confined to the reference pathway; it does not imply that the full transformer state is cached.

C.2 What Residual Conditioning Establishes

Residual bands respond to contours, texture, shadows, and illumination changes. Their normalization and pooling retain a compact summary rather than lossless boundary information. The transformer can use target context to interpret these cues, but successful segmentation does not prove exact restoration of VAE-discarded frequencies. Likewise, the results for the tested refiners do not imply that every post-decoding method must fail; a refiner with additional RGB information can access evidence beyond a compressed mask latent.

C.3 Decoded Segmentation Objectives

The interface $\mathcal{G}_s$ includes unpacking, VAE decoding at output resolution s, channel reduction, and probability conversion. We decode the endpoint latent once at 512×512 for the primary mask loss and separately at the higher resolution s_{hi} for the auxiliary loss. The two losses do not resize a single shared decoded output. Both decoder evaluations retain gradients to the same endpoint latent while decoder weights remain frozen. For aligned probabilities $p_i \in (0, 1)$ and targets $m_i \in [0, 1]$, the BCE and soft Dice objectives have the form

$$\ell_{\mathrm{BCE}}(p, m) = -\frac{1}{n_s} \sum_{i=1}^{n_s} \big[m_i \log p_i + (1 - m_i) \log(1 - p_i) \big], \tag{19}$$

$$\ell_{\mathrm{Dice}}(p, m) = 1 - \frac{2 \sum_i p_i m_i + \epsilon}{\sum_i p_i + \sum_i m_i + \epsilon}, \qquad \ell_{\mathrm{seg}} = \ell_{\mathrm{BCE}} + \ell_{\mathrm{Dice}}, \tag{20}$$

with $\epsilon > 0$ for numerical stability. The primary grid is 512×512; the auxiliary grid is finer. The detail objective is additional segmentation supervision, not a boundary-only, Fourier, or adversarial loss. The factorial results below independently switch reference tokens and this auxiliary term, avoiding attribution of their combined gain to either alone.

D Implementation and Evaluation Scope

Backbone and conditioning interfaces. The 9B runs use native Qwen3-8B, with transformer hidden width 4096 and text-feature width 12288. The 4B runs use native Qwen3-4B, with corresponding widths 3072 and 7680 (Labs, 2025). Backbone size refers to the generative transformer,

Table 9: Training and inference configuration.

Item	Setting
Training data / resolution	DIS-TR; 1024×1024
Hardware / training time	Four NVIDIA H100 GPUs, 94 GB each; approximately five hours
Reported peak memory	83 GB
Optimizer / budget	AdamW; 10,000 updates; global batch size 32
Learning rate	5×10^{-5}; halved at 512, 2048, 4096, 8192
Augmentation	Random flips and 90° rotations
Training timestep	$\mathrm{Beta}(2.5, 1.0)$
LoRA	Rank 128 on attention and MLP projections
Prompt sampling	One of four descriptions per image per step; cached embeddings
LPDP	16 Fourier levels; coordinate scale 1; three-layer GELU MLP; zero point projection
Training clicks	Two foreground and two background; uniform without replacement
Prompt dropout	Text 0.1; clicks 0.1
Evaluation clicks	Four fixed simulated clicks; seed 42; cached coordinates
DRTI	Three residual bands; 8×8 pooling; 16 channels; 256 width-128 slots
Loss weights	Flow 0.1; primary mask 1.0; auxiliary detail 0.1
Inference	Native backbone schedule; four explicit Euler steps

excluding the text encoder, VAE, and caption-generation model. This distinction also applies to the comparison with FlowDIS's 12B FLUX.1-schnell backbone (Sargsyan & Navasardyan, 2026).

Table 10: Tensor interfaces for the 9B configuration at 1024×1024.

Quantity	Shape	Role
z_I, z_M, z_t	4096×128	Packed image/mask/flow tokens
$[z_t \parallel z_I]$	4096×256	Input to widened trainable projection
c_{emb}	$L_T \times 4096$	Projected frozen text features
c_{pt}	$K \times 4096$	Trainable point content
c_{ctx}	$(L_T + K) \times 4096$	Concatenated context sequence
P	$16 \times 8 \times 8$	Pooled residual descriptor
R	256×128	Learned reference slots before input projection
v_θ	4096×128	Velocity on primary tokens only

Trainable parameters and resources. At rank 128, OmniDIS-9B has 143.88M trainable parameters, approximately 1.60% of the nominal 9B backbone size. The count includes LoRA adapters, the point projection, PointEncoder, widened input projection, and DRTI encoder. The pretrained transformer weights, VAE, and Qwen3 encoder remain frozen. The point projection is included in this total and is not added again. Training takes approximately five hours in the stated setup, with a reported peak memory of 83 GB. Cross-method FPS values describe their respective operating points.

E Additional Module Ablations

E.1 LPDP Encoding

Table 11 compares point representations with DRTI retained. Removing spatial rotary identifiers reduces F^{ω}_{β} from 0.949 to 0.939, a 1.0-point drop; removing continuous coordinate features reduces it to 0.934, a 1.5-point drop. Structure scores decrease by 0.7 and 1.1 points, respectively. Both representations contribute: Fourier features encode continuous position in token content, while rotary identifiers expose spatial relations within attention. Their combination performs best across all four reported metrics. The study isolates coordinate content and spatial attention identifiers.

Table 11: **LPDP encoding** DIS-VD controls with fixed text/click inputs and DRTI.

Point encoding	F_β^ω	F_β^{mx}	S_α	$\mathcal{M}\downarrow$
Coordinate features + polarity; no spatial RoPE	0.939	0.963	0.957	0.014
Spatial RoPE + polarity; no coordinate features	0.934	0.959	0.953	0.015
Full LPDP	0.949	0.970	0.964	.012

E.2 Reference Conditioning versus Auxiliary Supervision

Table 12 varies reference conditioning and the auxiliary detail loss independently, with LPDP active throughout. Without either factor, $F_\beta^\omega = 0.880$. Enabling only the auxiliary loss raises it to 0.894, a 1.4-point gain. Enabling reference tokens without the auxiliary loss instead reaches 0.939, a 5.9-point gain, showing that extra decoded supervision alone does not account for DRTI's contribution.

Combining references and the auxiliary loss yields $F_\beta^\omega = 0.949$, $S_\alpha = 0.964$, and $\mathcal{M} = 0.012$. The auxiliary term adds 1.0 point in F_β^ω when references are present, while references add 5.5 points when the auxiliary term is already active. The joint gain is slightly subadditive: $0.949 - 0.939 - 0.894 + 0.880 = -0.004$. Both factors improve accuracy, with a smaller auxiliary-loss gain when reference conditioning is already present.

Table 12: **DRTI/loss control**: LPDP is active in every proposed run.

Reference condition	Detail loss	F_β^ω	F_β^{mx}	S_α	$\mathcal{M}\downarrow$
Absent	Off	0.880	0.923	0.910	0.015
Absent	On	0.894	0.932	0.921	0.014
DRTI	Off	0.939	0.960	0.956	0.013
DRTI	On	0.949	0.970	0.964	0.012

F Failure Cases

Figure 6 shows errors on text-heavy and diagram-like images. The predictions merge fine strokes and interior spaces into larger foreground regions, whereas the reference masks retain sparse lines and character structure. These examples expose a distinction between selecting an object-sized region and preserving disconnected graphic elements. Strong residual responses alone do not determine which strokes belong to the target, and compact conditioning can miss the organization of such detail. The examples are qualitative failures, not a separate document-segmentation benchmark.



Figure 6: **Failure cases on graphic content.** Each triplet contains the input, reference mask, and OmniDIS prediction. Fine strokes and interior spaces are merged into coarser regions.