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

GlassWizard: Harvesting Diffusion Priors for Glass Surface Detection

Wenxue Li[1*] Tian Ye[1*] Xinyu Xiong[2] Jinbin Bai[3] Feilong Tang[4] Wenxuan Song[1]
Zhaohu Xing[1] Lie Ju[4] Guanbin Li[2] Lei Zhu[1,5✉]
[1]The Hong Kong University of Science and Technology (Guangzhou)
[2]Sun Yat-sen University [3]National University of Singapore [4]Monash University
[5]The Hong Kong University of Science and Technology

Abstract

Glass Surface Detection (GSD) is a critical task in computer vision, enabling precise interactions with transparent surfaces and enhancing both safety and object recognition accuracy. However, current research still faces challenges in both recognition performance and generalization capability. Thanks to the recent advanced diffusion-based generative models, GSD task can benefit from rich prior knowledge encapsulated in pre-trained Stable Diffusion (SD) model. Thus, in this paper, we present GlassWizard, aiming to harvest priors in diffusion-based model to achieve accurate and generalized GSD. Firstly, we delve into the text embedding space in SD to build an text-based context prior, thereby enhancing the understanding of implicit attribute of glass and achieving fine-grained predictions. Secondly, we train an end-to-end diffusion model with a one-step formulation pipeline, yielding effective optimization and fast inference. In addition, to facilitate our adapted framework scalable to other multi-modal GSD tasks (such as RGB-D/RGB-T GSD), we present a modality-customized adaptation, that can achieve rapid adaptation to multi-modal GSD tasks. Our experimental results demonstrate that our proposed framework achieves cutting-edge performance across diverse datasets, and it also shows strong generalization ability. Additionally, it excels in multi-modal GSD tasks, confirming its scalability across different modalities.

1. Introduction

Glass Surface Detection (GSD) [20, 25, 50, 51] is crucial for various applications, including robotics, autonomous driving, and augmented reality, where precise identification greatly enhances safety and interactivity in complex environments. Recent efforts have proposed many ingenious strategies for glass detection, such as exploring boundary information [8, 27], contextual features [54], and visual blurriness cues [29]. However, constructing an effective, unified model for glass detection remains an open challenge. Existing models often rely on small training datasets and have limited feature capacity, which restricts their ability to generalize across diverse real-world conditions.

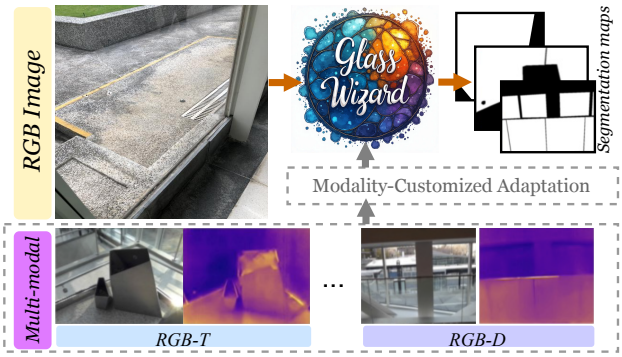

Figure 1. Our proposed GlassWizard harvests priors from diffusion-based model to achieve accurate and generalized Glass Surface Detection (GSD). It is also scalable to multi-modal data with Modality-Customized Adaptation.

Fortunately, with the construction of massive high-quality training datasets and well-designed large-scale model architectures, many computer vision tasks, such as image segmentation [17, 41] and image generation [32], have taken a leap forward in generalization and zero-shot capabilities in recent years. A prominent example of such models is the Segment Anything Model (SAM) [17], which demonstrates excellent zero-shot performance on many downstream segmentation tasks. Nevertheless, studies have shown that the standard SAM performs poorly on glass detection [7]. A key reason for this is that unlike explicit semantic classes (*e.g.*, person or car) that can be distinguished independently of context, accurately recognizing glass surfaces is highly context-dependent. To incorporate richer contextual semantics, some approaches introduce additional

*Equal contribution.
✉Lei Zhu is the corresponding author of this work.

modalities, such as paired depth [38] or thermal data [13] to improve segmentation accuracy. However, these approaches typically focus on architectural modifications tailored to individual modalities, integrating information from a single sensor type. There is significant value in developing frameworks that can scale to multiple modalities, accommodating a variety of sensor inputs. To our knowledge, a unified foundational model with multi-modal glass recognition capabilities has yet to be developed.

To address the aforementioned challenges, we turn our attention to Stable Diffusion (SD) [32], a large-scale image generation model that inherently supports flexible prompt conditioning like text and semantic map. The vast knowledge encapsulated in SD has been successfully transferred to various downstream tasks such as image classification [3, 36] and depth estimation [16]. Motivated by this, we explore the potential of leveraging the rich priors embedded in diffusion models to improve both the accuracy and generalization of glass surface detection.

Despite the impressive capabilities of SD, directly applying the SD framework to the GSD task presents several challenges. **First**, current GSD methods primarily rely on visual cues and often overlook the integration of additional sources of prior knowledge. In SD, the text embedding serves as a pivotal intermediary between text and images but remains relatively under-explored in the context of dense prediction tasks. **Second**, the VAE decoder of diffusion models, typically designed for image generation, may not align well with segmentation tasks that require a customized output mask. **Third**, the computational cost of multi-step inference in diffusion models is significantly higher than that of one-step segmentation models, which poses a challenge for efficient processing, particularly in real-time applications. **Finally**, while the stochastic multi-step generation process in SD allows the model to explore its generative capabilities and produce diverse outputs, it introduces noise that can lead to unwanted artifacts in the GSD task. In this context, the added noise in the latent space is unnecessary and may hinder the precise segmentation required for accurate glass surface detection, where fine-grained details are crucial.

To mitigate the aforementioned challenges, we propose GlassWizard, a framework that harnesses the potential of pre-trained diffusion-based models to improve performance and generalization ability for GSD, as demonstrated in Figure. 1. To effectively adapt Stable Diffusion (SD) for the GSD task, we propose an end-to-end training approach with an one-step formulation pipeline. On the one hand, the one-step approach reduces the randomness in the diffusion process, leading to more consistent and accurate outputs. On the other hand, it improves the efficiency of both training and inference, significantly reducing the computational cost and time required for the task. Meanwhile, we delve into the rich priors encapsulated in the textual embedding space of pre-trained SD and build text-based content prior for GSD. We refine the model's understanding of glass within the textual space by introducing a learnable textual condition representing the concept of glass surfaces. Additionally, we introduce a modality-customized adaptation, which effectively facilitates the scalability of the trained framework to other multi-modal tasks, such as RGB-D and RGB-T GSD. This enables the model to integrate features from different sensor modalities, making it adaptable to a wider range of real-world applications.

The main contributions are summarized as follows:

- We introduce GlassWizard, a general framework for glass surface detection that utilizes pre-trained text-to-image diffusion models. We employ an end-to-end training manner combined with a one-step formulation to enhance both efficiency and segmentation accuracy.
- We build the text-based content prior to explore and refine the potential of textual embeddings for learning the specific concepts of glass surfaces.
- We present a modality-customized adaptation for adapting the trained GSD framework to multi-modal GSD tasks, ensuring the model's scalability across different sensor suites.
- Extensive experiments demonstrate that GlassWizard not only outperforms state-of-the-art methods, but also shows excellent generalization when tested on two unseen GSD datasets.

2. Related Work

2.1. Transparent Image Segmentation

Transparent object segmentation has been a challenging task in computer vision due to the inherent properties of transparent materials, which often lack distinct texture and color information. To address the challenge, various approaches have been proposed, leveraging advanced deep learning techniques [8, 27, 29, 46, 47, 54, 55]. Some works explore boundary information [8, 27, 46], different-level feature fusion [54] Trans4Trans [55] presents a lightweight model to perform real-time way-finding in wearable system. VBNet [29] proposes to utilize visual blurriness information to detect glass. GlassSemNet [21] learns the spatial and semantic correlations between objects. VGSD-Net [23] explores dynamic reflection information for video glass surface detection. Moreover, several works utilize additional modalities, such as depth [38], polarized light [26], and thermal [13] to provide supplementary cues for segmentation. Existing GSD methods are limited by small training datasets, hindering their real-world applicability. There's a need for robust models that deliver high-quality segmentation without extensive fine-tuning, enhancing their practical use across various applications.

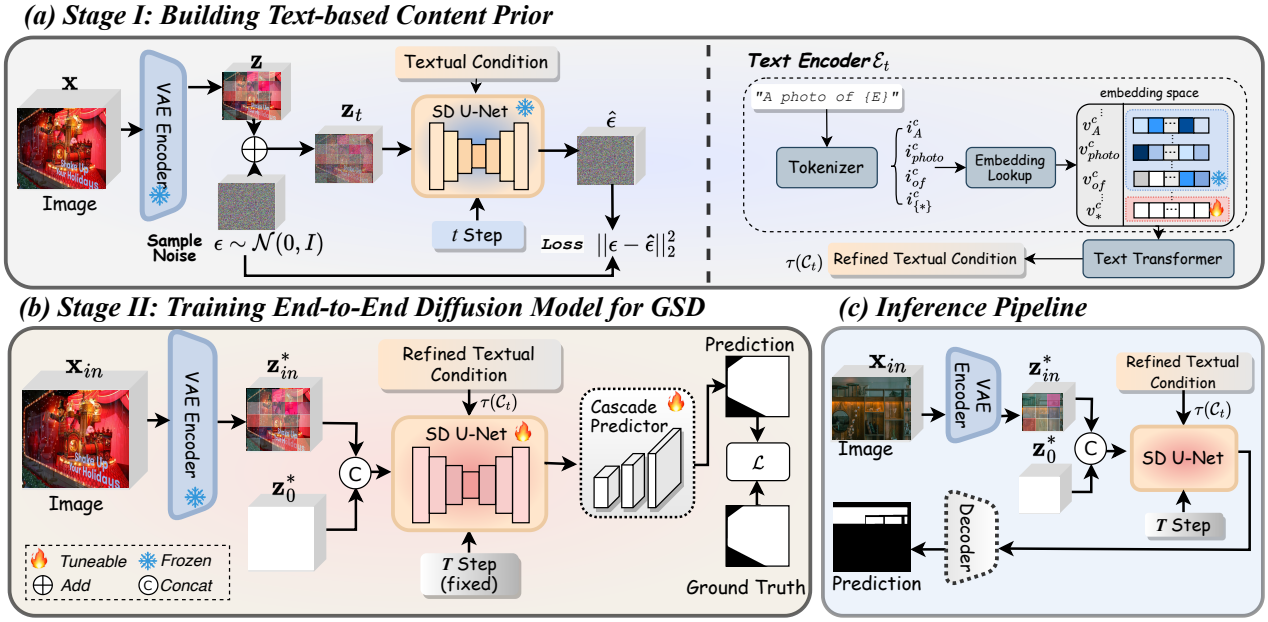

Figure 2. The overview of our proposed GlassWizard, which consists of two stages. (a) In *Stage I*, we train a refined textual condition of glass surface to build a text-based content prior for adapting SD. Most parameters of the pre-trained diffusion model are frozen, and only the glass-specific embedding is learned. (b) In *Stage II*, we train the diffusion model under the guidance of refined textual condition. The training is performed in an end-to-end manner with a one-step formulation pipeline. (c) Inference pipeline of our proposed method.

2.2. Diffusion Models

Diffusion models have recently gained prominence for their ability to generate high-quality images or videos [2, 28, 30, 33]. Key advancements include prominent models such as Denoising Diffusion Probabilistic Models (DDPMs) [10] and Latent Diffusion Model (LDM) [31], which reduce computational overhead by performing the diffusion process within a lower-dimensional latent space, significantly enhancing both training efficiency and inference speed. Conditional diffusion models incorporate various forms of conditional information, such as text prompt inputs [33], to steer the generation process towards specific desired outcomes. Stable Diffusion series [32] then extend LDM's approach, being impactful in enabling fine-grained control in image generation. The large-scale dataset, such as LAION-5B [34] allows the diffusion model to learn the complex relationships between image and text, and it enables diffusion models to generate high-quality images based on textual prompts, adjusting the generated content according to different descriptions. Therefore, text-to-image models possess rich prior knowledge that can be effectively harnessed.

2.3. Diffusion for Image Segmentation

Diffusion models can handle various downstream computer vision tasks [3, 5, 9, 16, 16, 18, 24, 36, 53], demonstrating their versatility and adaptability across various domains. SegDiff [1] is the first model that applies diffusion models to the image segmentation problem, which conditions the diffusion step estimation function on an input tensor that combines information derived from both the current estimate step and the input image. MedSegDiff [43] proposes a feature frequency parser, which allows for eliminate the negative effect of the high-frequency noise components in this process. Further, MedSegDiffV2 [44] proposes a Transformer-based Diffusion framework and a Neural Band-pass Filter (NBP-Filter) to align the noise and semantic features each time. DiffSeg [40] explores an unsupervised manner that leverages an iterative merging process to merge attention maps to generate segmentation masks. DMP [18] leverages pre-trained text-to-image diffusion models as a prior for dense prediction tasks. Despite the progress of existing diffusion model-based segmentation methods, effectively adapting pre-trained diffusion models for downstream tasks remains a challenge. Issues such as artifacts introduced by stochastic processes and high computational costs are prominent barriers.

3. Methodology

3.1. Preliminary: Latent Diffusion models

Diffusion models leverage diffusion process to synthesize desired high-quality images, which gradually transforms a

simple distribution (Gaussian noise) to a complex data distribution through a series denoising steps. In contrast to diffusion models that operate directly on the data, latent diffusion models (LDM) [31] execute diffusion steps within a low-dimensional latent space, thereby enhancing computational efficiency. It includes two-step process: forward nosing and reverse denoising. In the forward process, given an initial data sample $\mathbf{x}_0$, Gaussian noise is progressively added over T steps. At each step, the noisy data $\mathbf{x}_t$ is generated by:

$$\mathbf{x}_t = \sqrt{\alpha_t}\mathbf{x}_{t-1} + \sqrt{1-\alpha_t}\epsilon \tag{1}$$

where α_t is the variance schedule controlling the amount of noise added at each time step t and ϵ is Gaussian noise, $\epsilon \sim \mathcal{N}(0, I)$. The forward process can be represented as:

$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}). \tag{2}$$

The goal is to learn the reverse process, which gradually denoises $\mathbf{x}_t$ back to $\mathbf{x}_0$. This reverse process is modeled as:

$$p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t)), \tag{3}$$

where $\mu_\theta(\mathbf{x}_t, t)$ is the mean predicted by the model, and $\Sigma_\theta(\mathbf{x}_t, t)$ represents the variance. The ultimate goal is to learn the reverse process parameters θ so that the model can generate new samples starting from random noise.

3.2. Overall Architecture

In this work, we aim to adapt a pre-trained diffusion model for the GSD task, leveraging the rich prior knowledge embedded in the model to achieve generalized performance. As shown in Figure 2, our framework consists of two stages. In Stage I, we train a refined textual condition for glass surfaces, building a text-based content prior to represent the glass concept. Most parameters of the pre-trained diffusion model are frozen, and only the glass-specific embedding is learned. In Stage II, we adapt the Stable Diffusion model in an end-to-end manner, guided by the refined textual condition.

3.3. Stage I: Building Text-based Content Prior

Current studies for GSD predominantly rely on visual cues, often neglecting the integration of other domain knowledge, such as text. By leveraging large-scale datasets to align semantic features of text and visual elements, text-to-image diffusion models utilize text conditions and pre-trained language models to guide image synthesis, resulting in semantically coherent images. This raises a question: *Can we harness the potential of textual conditional priors in diffusion models for GSD task?* SD faces challenges in depicting glass surfaces due to their implicit semantics. The difficulty arises from the fact that glass objects lack distinct visual semantic features. To address this, we propose to learn a

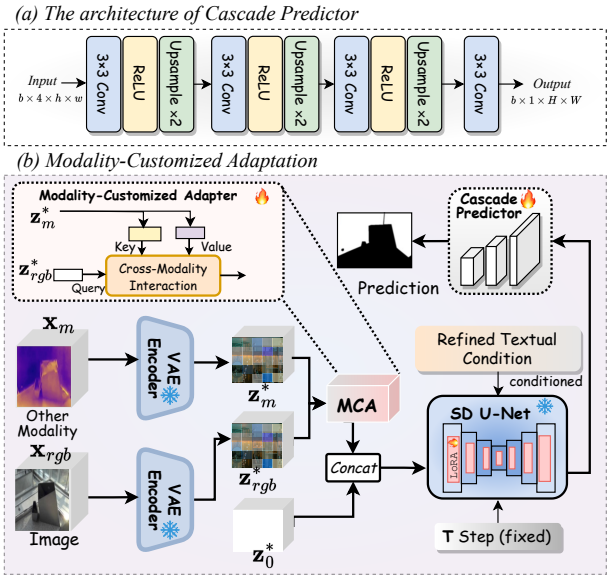

Figure 3. Modality-Customized Adaptation facilitates the trained diffusion model to integrate features from different sensor modalities, making it scalable to a wider range of real-world applications.

glass-specific text condition to represent the implicit features of glass surfaces.

We use the prompt template as "`A photo with [CLASS]`", where `[CLASS]` refers to the specific class. In our work, `[CLASS]` represents the concept of glass surfaces. To alleviate the above challenge that glass surface lacks clear semantic information, we define a learnable content in the prompt rather than `[CLASS]` to strengthen the concept of glass surfaces in the textual embedding space. The prompt can be formulated as

$$\mathcal{C}_t = \text{"A photo with [E}^*\text{]"}, \tag{4}$$

where `[E*]` is the placeholder of newly introduced word, representing the glass surface. We initialize `[E*]` with the word `glass`, allowing it to inherit some information associated with glass surface.

As shown in Figure 2 (a), we train the text-based content prior over Stable Diffusion model. The input image is encoded into a spatial latent code using the pre-trained VAE encoder as $\mathbf{z} = \mathcal{E}(\mathbf{x})$. At a randomly selected timestep t, the noised latent is denoted as $\mathbf{z}_t$. We freeze the other modules in SD, including the VAE encoder, SD U-Net, and text encoder, allowing only the newly introduced embedding vector to be trainable.

Textual Prompt Parameterization. The textual encoder $\mathcal{E}_t$ tokenizes the input text prompt $\mathcal{C}_t$ into a set of tokens, which is based on the index in its predefined dictionary, as $[i_1^c, \cdots, i_k^c]$ (k is the number of tokens). Note that we extend the original dictionary to represent the newly intro-

duced concept, which denotes the glass surface. Each token is linked to a specific embedding through an index lookup, represented as $[v_1^c, \cdots, v_k^c]$. However, the newly introduced word does not have a corresponding embedding token. Our objective is to optimize the embedding v_*^c for this new word.

Textual Condition Optimization. The noise estimation is denoted as $\hat{\epsilon} = \epsilon_\theta(\mathbf{x}, t, \tau(\mathcal{C}_t))$. We distill the semantic information of glass surface and get textual condition $\tau(\mathcal{C}_t)$ by minimizing the training objective with MSE loss, as

$$\mathcal{L}_{txt} = \mathbb{E}_{\mathbf{x}_0, \epsilon \sim \mathcal{N}(0,I), t \sim \mathcal{U}(T)} ||\epsilon - \hat{\epsilon}||_2^2. \tag{5}$$

The optimized embedding of newly introduced token `[E*]` is denoted as $v_*^c = \arg\min_v \mathbb{E}_{\mathbf{x}_0, \epsilon \sim \mathcal{N}(0,I), t \sim \mathcal{U}(T)} ||\epsilon - \hat{\epsilon}||_2^2$. This process could inject prior knowledge into the generative model, enhancing the model's adaptability to glass surface objects.

3.4. Stage II: Training End-to-End Diffusion Model for GSD

In Stage II, we adapt the Stable Diffusion model for GSD task, and the training pipeline is shown in Figure 2 (b). Given the input image $\mathbf{x}_{in} \in \mathbb{R}^{C \times H \times W}$ and its corresponding ground truth mask $\mathbf{m} \in \mathbb{R}^{H \times W}$, with the number of channels C, and spatial resolution $H \times W$, we first project the input image into the latent space using the VAE encoder $\mathcal{E}(\cdot)$, as $\mathbf{z}_{in}^* = \mathcal{E}(\mathbf{x}_{in})$. In diffusion models, the introduced stochastic noise is designed for generative diversity. The original learning protocol of SD that learns to predict noise, is not suitable for deterministic task. Thus, in order to avoid the negative impact of stochastic noise, we do not start from noisy latent. Instead, we initialize a zero latent vector, as $\mathbf{z}_0^* = \mathbf{0}$ and we concatenate the RGB latent $\mathbf{z}$ and the zero latent $\mathbf{z}_0^*$ as $\mathbf{z}_c = concat(\mathbf{z}_{in}^*, \mathbf{z}_0^*)$, where $\mathbf{z}_c$ is the input of the SD U-Net. We utilize the learned textual condition in conditional SD U-Net $f_\theta(\cdot)$ to as $f_\theta(\mathbf{x}_{in}, T, \tau(\mathcal{C}_t))$. Here, to prevent the activation magnitudes of the first layer from becoming too large and to retain the pre-trained structure, we duplicate the input layer's weights and halve its values.

Enabling One-step Sampling. Multi-step diffusion process is computation consumption. We condense the diffusion process into a single step, simplifying the optimization process, and significantly improving inference speed. Different with the stochastic multi-step generation, we fix $t = T$ to train the model for single-step prediction.

End-to-End Training. The primary training goal of the original diffusion model is denoising. However, it does not necessarily ensure strong performance for deterministic task. Focusing solely on denoising might introduce larger pixel variance, resulting in less stable predictions. To alleviate this issue, we introduce the end-to-end manner. Here, we introduce a simple but effective Cascade Predictor for adapting the GSD task, which is shown in Figure 3 (a). The output of SD U-Net is denoted as $\mathbf{z}' = f_\theta(\mathbf{x}_{in}, T, \tau(\mathcal{C}_t))$, then it was fed to Cascade Predictor to predict segmentation maps. In our framework, the model is trained to directly predict segmentation mask instead of noise. Following the approaches in [42], we use the weighted IoU loss and BCE loss as our training objectives. The output segmentation mask $\hat{\mathbf{m}}$ is the optimization target as

$$\mathcal{L} = \mathcal{L}_{IoU}^w(\mathbf{m}, \hat{\mathbf{m}}) + \mathcal{L}_{BCE}^w(\mathbf{m}, \hat{\mathbf{m}}). \tag{6}$$

Such a framework takes advantage of the diffusion model's natural strengths, allowing us to leverage extensive knowledge prior of high-quality natural images. This improves the model's generalization ability while also leading to a substantial gain in efficiency.

3.5. Inference: One-Step Pipeline

The prediction pipeline is illustrated in Figure 2 (c). Unlike the diffusion pipeline, which employs the VAE decoder to decode latent code and generate images, our framework aligns the training pipeline with the inference process. The timestep is also fixed $t = T$. This single-step inference significantly enhances inference efficiency compared with multi-step formulation in diffusion models.

3.6. Modality-Customized Adaptation

In this section, we focus on adapting the previously trained GSD model to accommodate additional modalities. The key challenge in multi-modal GSD tasks is effectively integrating the RGB modality with other modalities to enhance performance. The adaptation pipeline is shown in Figure 3.

Modality-Customized Adapter. We project the input RGB image $\mathbf{x}_{rgb}$ and modality-specific input $\mathbf{x}_m$ into the latent space using the VAE encoder, as $\mathbf{z}_{rgb}^* = \mathcal{E}(\mathbf{x}_{rgb})$, $\mathbf{z}_m^* = \mathcal{E}(\mathbf{x}_m)$. Then, we aggregate $\mathbf{z}_{rgb}^*$ and $\mathbf{z}_m^*$ using cross-attention operation to obtain the modality-specific activated feature $\mathbf{z}_a^*$ We take $\mathbf{z}_{rgb}^*$ as the query, and $\mathbf{z}_m^*$ as both key and value. The query, key and value are fed into projected layer, achieved by an 1×1 convolution operation, which can be written as: $\mathbf{z}_i^Q = \phi_q(\mathbf{z}_{rgb}^*), \mathbf{z}_i^K = \phi_k(\mathbf{z}_m^*), \mathbf{z}_i^V = \phi_v(\mathbf{z}_m^*)$, where $\phi_q(\cdot)$, $\phi_k(\cdot)$, $\phi_v(\cdot)$ denote the projected operation. Then, we employ these projected features to conduct cross-modality interaction, as:

$$\mathbf{z}_a^* = \mathbf{z}_{rgb}^* + \text{softmax}(\frac{\mathbf{z}_i^Q \cdot \mathbf{z}_i^{K^T}}{\sqrt{d}})\mathbf{z}_i^V, \tag{7}$$

where d is the dimension of the key vectors. We incorporate LoRA layers [11] in SD U-Net to facilitate rapid adaptation to various multi-modal GSD tasks without the risk of catastrophic forgetting.

4. Experiments

4.1. Datasets

Stage I: We combine the training sets of *GDD* [25], *Trans10K-Stuff* [46], *GSD* [20] and *HSO* [54] for building

Table 1. Model performance for GSD task on the GDD [25], Trans10K-Stuff [46], GSD [20] and HSO [54] datasets.

Model	GDD [25]				Trans10K-Stuff [25]				GSD [20]				HSO [54]			
	IoU↑	F_β↑	MAE↓	BER↓	IoU↑	F_β↑	MAE↓	BER↓	IoU↑	F_β↑	MAE↓	BER↓	IoU↑	F_β↑	MAE↓	BER↓
Trans2Seg [47]	0.854	0.929	0.076	7.14	0.869	0.935	0.057	5.69	0.781	0.880	0.072	9.47	0.764	0.875	0.102	10.57
GSDNet [20]	0.859	0.929	0.073	6.83	0.873	0.936	0.055	5.36	0.796	0.882	0.068	8.56	0.770	0.878	0.101	10.00
GDNet-B [27]	0.840	0.923	0.088	7.85	0.859	0.929	0.064	6.16	0.781	0.877	0.078	9.56	0.752	0.866	0.113	11.16
PGSNet [54]	0.880	0.932	0.074	6.95	0.889	0.944	0.049	5.65	0.850	0.906	0.062	5.98	0.823	0.921	0.068	7.09
VBNet [29]	0.893	0.948	0.063	5.43	0.878	0.935	0.059	5.00	0.853	0.916	0.059	5.89	0.815	0.905	0.090	8.15
DDP [14]	0.908	0.958	0.054	4.39	0.906	0.954	0.044	3.92	0.865	0.927	0.053	5.38	0.839	0.926	0.075	7.07
Marigold [16]	0.731	0.821	0.201	16.88	0.739	0.837	0.163	14.02	0.658	0.777	0.191	17.97	0.580	0.707	0.291	24.75
Ours	**0.933**	**0.969**	**0.039**	**3.62**	**0.928**	**0.965**	**0.030**	**3.04**	**0.904**	**0.952**	**0.036**	**4.03**	**0.879**	**0.941**	**0.055**	**5.44**

Table 2. Zero-shot performance of our proposed method for video glass surface detection on the VGSD-D dataset [23]. The compared methods are trained on the training set of VGSD-D dataset.

Model	IoU↑	MAE↓	BER↓
GDNet [25]	0.735	0.172	13.18
EBLNet [8]	0.764	0.134	13.25
PGSNet [54]	0.703	0.156	15.11
SC-Cor [4]	0.765	0.125	12.15
UFO [37]	0.634	0.254	22.43
VMD [22]	0.763	0.123	12.44
VGSD-Net [23]	0.802	0.099	9.54
Ours	**0.942**	**0.031**	**2.92**

Table 3. Zero-shot performance of our proposed method on RGBP dataset [26]. Here, we only adopt the RGB iamge as the input to test zero-shot segmentation ability of the model. The compared methods are trained on the training set of RGBP dataset.

Model	IoU↑	MAE↓	BER↓
GDNet [25]	0.776	0.119	11.79
GSDNet [20]	0.781	0.122	12.61
EAFNet [45]	0.539	0.125	12.15
P Mask R-CNN [15]	0.660	0.178	18.92
PGSNet-Polar [26]	**0.811**	**0.091**	**9.63**
Ours	0.792	0.108	11.18

the textual content prior in Stage I.

Stage II: To train our framework, we utilize the combination of the training sets of *GDD* [25], *Trans10K-Stuff* [46] [54], *GSD* [20] and *HSO* [54]. WWe test our model on the test sets of these datasets to assess learning efficiency. To evaluate the generalization ability of our method, we perform zero-shot segmentation using the trained model on the test sets of *VGSD-D* [23] and *RGBP* [26] datasets. Here, we only adopt the RGB images in *RGBP*.

Modality-Customized Adaptation: We test the modality-customized adaptation performance on RGB-D and RGB-T GSD tasks. For RGB-D GSD task, we adopt *TROSD* dataset [38]. For RGB-T GSD task, we use *RGBT* [13] dataset.

Table 4. Ablation studies for GSD task on the GDD [25] and GSD [20] datasets.

Component	GDD [25]				GSD [20]			
	IoU↑	F_β↑	MAE↓	BER↓	IoU↑	F_β↑	MAE↓	BER↓
SD Adaptation	0.713	0.805	0.209	18.56	0.658	0.777	0.191	17.97
w/o CP	0.866	0.922	0.109	8.56	0.778	0.865	0.112	14.70
w/o RTC	0.923	0.956	0.057	3.67	0.890	0.939	0.040	4.26
TAE Encoder	0.929	0.969	0.038	3.60	0.896	0.948	0.035	4.38
Gaussian noise	0.919	0.958	0.045	3.89	0.889	0.938	0.042	4.30
Ours	**0.933**	**0.969**	**0.039**	**3.62**	**0.904**	**0.952**	**0.036**	**4.03**

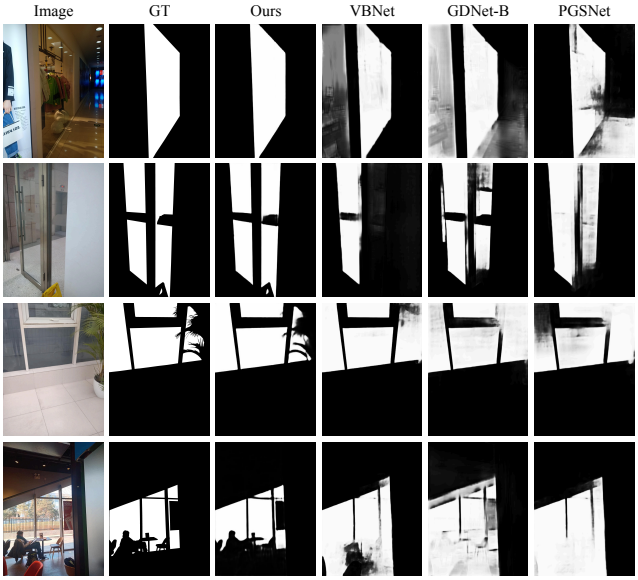

Figure 4. Qualitative comparison of different methods.

4.2. Experimental Setup

Implementation Details. We resize the images into 512×512 in both training and evaluation. We adopt the Stable Diffusion 2.0 as our base model. In Stage I, we train the text embeddings for 10 epochs with AdamW optimizer. The learning rate is set to 5e-04. The batch size is 10. In Stage II, the learning rate is initialized to 3e-05 and is scheduled by LambdaLR. The number of training epochs is set to 20, and the batch size is set to 3. The model is trained with Adam Optimizer. The timestep T is set to 999. We use random horizontal/vertical flipping for data augmentation. Our method is implemented using PyTorch and trained on a single NVIDIA RTX 4090 GPU with 24GB of memory.

Evaluation Metrics. Following previous methods [54], we adopt Intersection over Union (IoU), weighted F-measure metric (F_β), mean absolute error (MAE), and balance error rate (BER) to evaluate the segmentation performance.

4.3. Comparison Studies

Quantitative Comparison. We choose state-of-the-art GSD methods, including Trans2Seg [47], GSDNet [20], GDNet-B [27], PGSNet [54], VBNet [29] for comparison. We also adopt diffusion-based methods, DDP [14], DMP [18] and Marigold [16] for comparison. For fair comparison, we re-implement these methods with our training pipeline. As shown in Table 1, our method outperforms other compared methods across all datasets, indicating that our method is superior for GSD task. Overall, the experimental results demonstrate that the knowledge encapsulated in Stable Diffusion are beneficial for GSD task. Moreover, we train our framework on the dataset separately to prove the effectiveness on the single dataset, and our method also achieve state-of-the-art performance. (see Supplementary Material).

Qualitative Comparison. In Figure 4, we present a qualitative comparison of different methods. Benefiting from the powerful capabilities of Stable Diffusion and learned text-based content prior, our model captures more semantic information of glass surface structures, thereby achieving improved localization of glass surface. Moreover, diffusion model can also capture fine-grained details, providing fine-grained texture information for GSD.

Generalization ability. We conducted tests on two previously unseen datasets, VGSD-D [23] and RGBP [26]. As shown in Table 2 and Table 3. The results of our model outperforms all methods trained on the VGSD-D training set, demonstrating the generalization ability and practical usability of our model. For RGBP dataset, we only adopt RGB modality without extra modality information. Our model achieved results comparable to the state-of-the-art that utilizes the multi-modal training set. This shows that our model achieves strong performance and can be a out-of-box model without the need for fine-tuning.

Table 5. Ablation studies for model efficiency. T denotes the timestep of the diffusion process, and Nums refers to the number of ensembled multi-step outputs. T_{inf} is the inference time required to generate an image with a resolution of 512, and it is measured using an NVIDIA RTX 4090 GPU.

Method	T	Nums	T_{inf}	GDD [25]			
				IoU↑	F_β↑	MAE↓	BER↓
Marigold	50	10	14.03s	0.731	0.821	0.201	16.88
Marigold	50	1	2.12s	0.713	0.805	0.209	18.56
Marigold	1	1	0.215s	0.284	0.530	0.496	49.99
Ours	1	1	0.177s	**0.933**	**0.969**	**0.039**	**3.62**

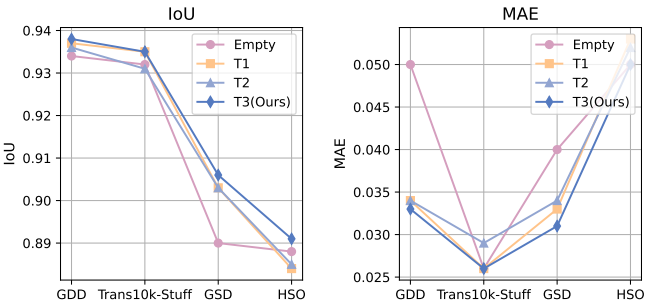

Figure 5. Model performance under different text conditions on different datasets. We use the prompt format "A photo with [E*]" where [E*] represents the textual condition. In T1, [E*]="transparent objects"; in T2, [E*]="glass"; and in T3, [E*] is the refined textual condition explicitly learned from glass surface images..

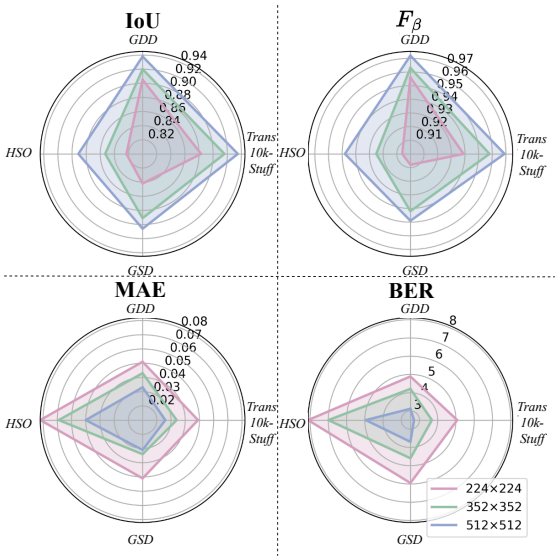

Figure 6. Model performance under different input resolution.

4.4. Ablation Studies

As shown in Table 4, we conduct ablation studies to evaluate the contribution of each component in our framework. First, we adapt Stable Diffusion using a two-stage pipeline. The model is optimized with the standard diffusion objective and inferred using multi-step DDIM (denoted as SD Adaptation). Next, we replace the Cascade Predictor with a VAE decoder (denoted as w/o CP) and conduct experiments. Adapting SD without the Cascade Predictor leads to a drop in performance, highlighting the importance of using an appropriate head for the GSD task. We also investigate the impact of removing the refined textual condition (RTC) and Stage I training separately. The results demonstrate that the textual condition is critical for training, and incorporating the trained textual condition leads to significant improvements in performance. Moreover, we replaced the VAE encoder with a more lightweight TAE encoder. However, the performance of the TAE encoder was inferior to that of the VAE encoder. To validate the effectiveness of the added all-zero latent instead of random noise, we replaced $\mathbf{z}_0^*$ with Gaussian noise. The results confirms the effectiveness of our framework in deterministic model.

Table 6. Model performance for RGB-D GSD task on the TROSD dataset [38].

Model	Input	IoU↑
TransLab [46]	RGB	0.507
Trans4Trans [55]	RGB	0.392
Ours w/o MCA	RGB	**0.628**
EBLNet [8]	RGB-D	0.501
EMSANet [35]	RGB-D	0.441
TROSNet [38]	RGB-D	0.572
Ours	RGB-D	**0.667**

Table 7. Model performance for RGB-T GSD task on the RGBT dataset [13].

Model	Input	MAE↓
Segformer [48]	RGB	0.053
EBLNet [55]	RGB	0.104
RGBT [38] (RGB)	RGB	0.052
Ours w/o MCA	RGB	**0.034**
Zhang *et al*. [56]	RGB-T	0.163
RGBT [38]	RGB-T	0.024
Ours	RGB-T	**0.022**

Modality-Customized Adaptation. We compare the multi-modal adaptation performance with state-of-the-art RGB, RGB-D and RGB-T GSD methods as shown in Table 6 and Table 7. The experimental results demonstrate the effectiveness of our strategy of Modality-Customized Adaptation (MCA), which fuse the extra modality. Meanwhile, even when the input consists solely of the RGB modality, we still achieved better performance, further proving the effectiveness of the framework.

4.5. Further Analysis

Model Efficiency. As shown in Table 5, Marigold incurs high costs due to its multi-step diffusion process. Our proposed method demonstrates fast inference times without sacrificing performance.

How Different Textual Conditions Affect Performance? We explore the performance with different input conditions, and the results are shown in Figure 5. When the textual condition is empty, performance is at its lowest compared to when it is utilized, highlighting the fact that the textual condition is crucial for the GSD task. Compared with the prompt word of `"glass"` or `"transparent objects"`, our proposed framework obtains better results when using our learned textual condition.

Input Resolution. We analyze a set of input resolutions to our approach, including 352×353, 512×512 and 768×768 with the results presented in Figure 6. Generally, a larger resolution typically results in better performance.

Transferability on Mirror Detection Task. We perform experiments on the mirror detection task and we employ the PMD [19] and MSD [52] datasets. We train the model for Stage I and Stage II on the PMD and MSD datasets separately. The evaluation was carried out using the standard test splits of each dataset. The results are shown in Table 8. Our method consistently outperforms other state-of-the-art mirror detection methods [6, 12, 39, 49] by a large margin, indicating that our proposed method is superior in transferability on other implicit object segmentation tasks.

Table 8. Quantitative comparison for mirror detection task on PMD [19] and MSD [52] datasets.

Method	PMD [19]				MSD [52]			
	IoU↑	F_β↑	MAE↓	BER↓	IoU↑	F_β↑	MAE↓	BER↓
SANet [6]	0.668	0.795	0.071	13.44	0.799	0.877	0.054	8.31
VCNet [39]	0.640	0.811	0.035	15.68	0.801	0.897	0.048	8.43
SATNet [12]	0.694	0.847	0.025	11.93	0.854	0.922	0.033	6.21
CSFwinformer [49]	0.701	0.838	0.024	11.41	82.08	0.896	0.045	7.14
Ours	**0.778**	**0.857**	**0.022**	**6.88**	**0.911**	**0.950**	**0.023**	**3.27**

5. Conclusion

In this paper, we propose GlassWizard, a framework that aims at harvesting priors from the diffusion-based model to achieve accurate and generalized Glass Surface Detection (GSD). First, we explore the textual embedding space in Stable Diffusion (SD) to construct a text-based contextual prior, which enhances the model's understanding of implicit glass attributes. Second, we introduce an efficient end-to-end training pipeline that enables effective adaptation to downstream GSD tasks in a single-step formulation. Additionally, we introduce a modality-customized adaptation strategy, which facilitates rapid adaptation across various multi-modal GSD tasks, such as RGB-D and RGB-T GSD. Our experimental results demonstrate that our proposed method delivers state-of-the-art performance across multiple benchmarks, and it also exhibits strong generalization capabilities. Moreover, after adapting for multi-modal GSD datasets, our framework excels in multi-modal GSD tasks, highlighting its scalability to different modalities.

Acknowledgments

This work is supported by the Guangdong Science and Technology Department (No. 2024ZDZX2004), the China Southern Power Grid Science and Technology Project (Project No.: 030117KC23120005), the Nansha Key Area Science and Technology Project (No. 2023ZD003), Guangzhou-HKUST(GZ) Joint Funding Program (No. 2024A03J0618), and Guangdong Provincial Key Lab of Integrated Communication, Sensing and Computation for Ubiquitous Internet of Things(No.2023B1212010007).

References

[1] Tomer Amit, Tal Shaharbany, Eliya Nachmani, and Lior Wolf. Segdiff: Image segmentation with diffusion probabilistic models. *arXiv preprint arXiv:2112.00390*, 2021. 3

[2] Dreamlike Art. Dreamlike photoreal 2.0. https://huggingface.co/dreamlike-art/dreamlike-photoreal-2.0, 2023. 3

[3] Kevin Clark and Priyank Jaini. Text-to-image diffusion models are zero shot classifiers. In *NeurIPS*, 2024. 2, 3

[4] Xinpeng Ding, Jingwen Yang, Xiaowei Hu, and Xiaomeng Li. Learning shadow correspondence for video shadow detection. In *ECCV*, pages 705–722. Springer, 2022. 6

[5] Gonzalo Martin Garcia, Karim Abou Zeid, Christian Schmidt, Daan de Geus, Alexander Hermans, and Bastian Leibe. Fine-tuning image-conditional diffusion models is easier than you think. *arXiv preprint arXiv:2409.11355*, 2024. 3

[6] Huankang Guan, Jiaying Lin, and Rynson WH Lau. Learning semantic associations for mirror detection. In *CVPR*, pages 5941–5950, 2022. 8

[7] Dongsheng Han, Chaoning Zhang, Yu Qiao, Maryam Qamar, Yuna Jung, SeungKyu Lee, Sung-Ho Bae, and Choong Seon Hong. Segment anything model (sam) meets glass: Mirror and transparent objects cannot be easily detected. *arXiv preprint arXiv:2305.00278*, 2023. 1

[8] Hao He, Xiangtai Li, Guangliang Cheng, Jianping Shi, Yunhai Tong, Gaofeng Meng, Véronique Prinet, and LuBin Weng. Enhanced boundary learning for glass-like object segmentation. In *ICCV*, pages 15859–15868, 2021. 1, 2, 6, 8

[9] Jing He, Haodong Li, Wei Yin, Yixun Liang, Leheng Li, Kaiqiang Zhou, Hongbo Liu, Bingbing Liu, and Ying-Cong Chen. Lotus: Diffusion-based visual foundation model for high-quality dense prediction. *arXiv preprint arXiv:2409.18124*, 2024. 3

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020. 3

[11] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022. 5

[12] Tianyu Huang, Bowen Dong, Jiaying Lin, Xiaohui Liu, Rynson WH Lau, and Wangmeng Zuo. Symmetry-aware transformer-based mirror detection. In *AAAI*, pages 935–943, 2023. 8

[13] Dong Huo, Jian Wang, Yiming Qian, and Yee-Hong Yang. Glass segmentation with rgb-thermal image pairs. *IEEE Transactions on Image Processing*, 32:1911–1926, 2023. 2, 6, 8

[14] Yuanfeng Ji, Zhe Chen, Enze Xie, Lanqing Hong, Xihui Liu, Zhaoqiang Liu, Tong Lu, Zhenguo Li, and Ping Luo. Ddp: Diffusion model for dense visual prediction. In *ICCV*, pages 21741–21752, 2023. 6, 7

[15] Agastya Kalra, Vage Taamazyan, Supreeth Krishna Rao, Kartik Venkataraman, Ramesh Raskar, and Achuta Kadambi. Deep polarization cues for transparent object segmentation. In *CVPR*, pages 8599–8608, 2020. 6

[16] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *CVPR*, 2024. 2, 3, 6, 7

[17] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, pages 4015–4026, 2023. 1

[18] Hsin-Ying Lee, Hung-Yu Tseng, and Ming-Hsuan Yang. Exploiting diffusion prior for generalizable dense prediction. In *CVPR*, pages 7861–7871, 2024. 3, 7

[19] Jiaying Lin, Guodong Wang, and Rynson WH Lau. Progressive mirror detection. In *CVPR*, pages 3697–3705, 2020. 8

[20] Jiaying Lin, Zebang He, and Rynson WH Lau. Rich context aggregation with reflection prior for glass surface detection. In *CVPR*, pages 13415–13424, 2021. 1, 5, 6, 7

[21] Jiaying Lin, Yuen-Hei Yeung, and Rynson W.H. Lau. Exploiting semantic relations for glass surface detection. In *NeurIPS*, 2022. 2

[22] Jiaying Lin, Xin Tan, and Rynson WH Lau. Learning to detect mirrors from videos via dual correspondences. In *CVPR*, pages 9109–9118, 2023. 6

[23] Fang Liu, Yuhao Liu, Jiaying Lin, Ke Xu, and Rynson WH Lau. Multi-view dynamic reflection prior for video glass surface detection. In *AAAI*, pages 3594–3602, 2024. 2, 6, 7

[24] Zexiang Liu, Yangguang Li, Youtian Lin, Xin Yu, Sida Peng, Yan-Pei Cao, Xiaojuan Qi, Xiaoshui Huang, Ding Liang, and Wanli Ouyang. Unidream: Unifying diffusion priors for relightable text-to-3d generation. In *ECCV*, pages 74–91. Springer, 2025. 3

[25] Haiyang Mei, Xin Yang, Yang Wang, Yuanyuan Liu, Shengfeng He, Qiang Zhang, Xiaopeng Wei, and Rynson W.H. Lau. Don't hit me! glass detection in real-world scenes. In *CVPR*, 2020. 1, 5, 6, 7

[26] Haiyang Mei, Bo Dong, Wen Dong, Jiaxi Yang, Seung-Hwan Baek, Felix Heide, Pieter Peers, Xiaopeng Wei, and Xin Yang. Glass segmentation using intensity and spectral polarization cues. In *CVPR*, pages 12622–12631, 2022. 2, 6, 7

[27] Haiyang Mei, Xin Yang, Letian Yu, Qiang Zhang, Xiaopeng Wei, and Rynson WH Lau. Large-field contextual feature learning for glass detection. *IEEE TPAMI*, 45(3):3329–3346, 2023. 1, 2, 6, 7

[28] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and

Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *ICLR*, 2024. 3

[29] Fulin Qi, Xin Tan, Zhizhong Zhang, Mingang Chen, Yuan Xie, and Lizhuang Ma. Glass makes blurs: Learning the visual blurriness for glass surface detection. *IEEE Transactions on Industrial Informatics*, 2024. 1, 2, 6, 7

[30] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 3

[31] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 3, 4

[32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 1, 2, 3

[33] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, pages 36479–36494, 2022. 3

[34] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *NeurIPS*, 35:25278–25294, 2022. 3

[35] Daniel Seichter, Söhnke Benedikt Fischedick, Mona Köhler, and Horst-Michael Groß. Efficient multi-task rgb-d scene analysis for indoor environments. In *2022 International joint conference on neural networks (IJCNN)*, pages 1–10. IEEE, 2022. 8

[36] Sitian Shen, Zilin Zhu, Linqian Fan, Harry Zhang, and Xinxiao Wu. Diffclip: Leveraging stable diffusion for language grounded 3d classification. In *WACV*, pages 3596–3605, 2024. 2, 3

[37] Yukun Su, Jingliang Deng, Ruizhou Sun, Guosheng Lin, Hanjing Su, and Qingyao Wu. A unified transformer framework for group-based segmentation: Co-segmentation, co-saliency detection and video salient object detection. *IEEE TMM*, 26:313–325, 2023. 6

[38] Tianyu Sun, Guodong Zhang, Wenming Yang, Jing-Hao Xue, and Guijin Wang. Trosd: A new rgb-d dataset for transparent and reflective object segmentation in practice. *IEEE TCSVT*, 33(10):5721–5733, 2023. 2, 6, 8

[39] Xin Tan, Jiaying Lin, Ke Xu, Pan Chen, Lizhuang Ma, and Rynson WH Lau. Mirror detection with the visual chirality cue. *IEEE TPAMI*, 45(3):3492–3504, 2022. 8

[40] Junjiao Tian, Lavisha Aggarwal, Andrea Colaco, Zsolt Kira, and Mar Gonzalez-Franco. Diffuse attend and segment: Unsupervised zero-shot segmentation using stable diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3554–3563, 2024. 3

[41] Hongqiu Wang, Guang Yang, Shichen Zhang, Jing Qin, Yike Guo, Bo Xu, Yueming Jin, and Lei Zhu. Video-instrument synergistic network for referring video instrument segmentation in robotic surgery. *IEEE Transactions on Medical Imaging*, 2024. 1

[42] Jun Wei, Shuhui Wang, and Qingming Huang. F^3net: fusion, feedback and focus for salient object detection. In *AAAI*, pages 12321–12328, 2020. 5

[43] Junde Wu, Rao Fu, Huihui Fang, Yu Zhang, Yehui Yang, Haoyi Xiong, Huiying Liu, and Yanwu Xu. Medsegdiff: Medical image segmentation with diffusion probabilistic model. In *Medical Imaging with Deep Learning*, pages 1623–1639. PMLR, 2024. 3

[44] Junde Wu, Wei Ji, Huazhu Fu, Min Xu, Yueming Jin, and Yanwu Xu. Medsegdiff-v2: Diffusion-based medical image segmentation with transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6030–6038, 2024. 3

[45] Kaite Xiang, Kailun Yang, and Kaiwei Wang. Polarization-driven semantic segmentation via efficient attention-bridged fusion. *Opt. Express*, 29(4):4802–4820, 2021. 6

[46] Enze Xie, Wenjia Wang, Wenhai Wang, Mingyu Ding, Chunhua Shen, and Ping Luo. Segmenting transparent objects in the wild. In *ECCV*, pages 696–711. Springer, 2020. 2, 5, 6, 8

[47] Enze Xie, Wenjia Wang, Wenhai Wang, Peize Sun, Hang Xu, Ding Liang, and Ping Luo. Segmenting transparent object in the wild with transformer. In *IJCAI*, 2021. 2, 6, 7

[48] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *NeurIPS*, 2021. 8

[49] Zhifeng Xie, Sen Wang, Qiucheng Yu, Xin Tan, and Yuan Xie. Csfwinformer: Cross-space-frequency window transformer for mirror detection. *IEEE TIP*, 2024. 8

[50] Zhaohu Xing, Lihao Liu, Tian Ye, Sixiang Chen, Yijun Yang, Guang Liu, Xiaojie Xu, and Lei Zhu. Farther than mirror: Explore pattern-compensated depth of mirror with temporal changes for video mirror detection. 1

[51] Zhaohu Xing, Lihao Liu, Yijun Yang, Hongqiu Wang, Tian Ye, Sixiang Chen, Wenxue Li, Guang Liu, and Lei Zhu. Detect any mirrors: Boosting learning reliability on large-scale unlabeled data with an iterative data engine. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25476–25486, 2025. 1

[52] Xin Yang, Haiyang Mei, Ke Xu, Xiaopeng Wei, Baocai Yin, and Rynson WH Lau. Where is my mirror? In *ICCV*, pages 8809–8818, 2019. 8

[53] Tian Ye, Sixiang Chen, Wenhao Chai, Zhaohu Xing, Jing Qin, Ge Lin, and Lei Zhu. Learning diffusion texture priors for image restoration. In *CVPR*, pages 2524–2534, 2024. 3

[54] Letian Yu, Haiyang Mei, Wen Dong, Ziqi Wei, Li Zhu, Yuxin Wang, and Xin Yang. Progressive glass segmentation. *IEEE TIP*, 2022. 1, 2, 5, 6, 7

[55] Jiaming Zhang, Kailun Yang, Angela Constantinescu, Kunyu Peng, Karin Müller, and Rainer Stiefelhagen. Trans4trans: Efficient transformer for transparent object segmentation to

help visually impaired people navigate in the real world. In *ICCV*, pages 1760–1770, 2021. 2, 8

[56] Qiang Zhang, Tonglin Xiao, Nianchang Huang, Dingwen Zhang, and Jungong Han. Revisiting feature fusion for rgb-t salient object detection. *IEEE TCSVT*, 31(5):1804–1818, 2021. 8