ST-SAM: Multimodal Scene Text Segmentation with Dense Visual and Sparse Textual Prompts via SAM

Jin Wei[1], **Yaqiang Wu**[1,2] *, **Jiayi Yan**[3], **Zeng Li**[4], **Zhen Xu**[1], **Yu Zhou**[5], **Lingling Zhang**[2], **QianYing Wang**[6]

[1] Lenovo AI Technology Center, CTOO, Lenovo
[2] School of Computer Science and Technology, Xi'an Jiaotong University
[3] Tsinghua University
[4] Institute of Information Engineering, Chinese Academy of Sciences
[5] VCIP & TMCC & DISSec, College of Computer Science & College of Cryptology and Cyber Science, Nankai University
[6] Technology Strategy & Platforms, CTOO, Lenovo
weijin4@lenovo.com, wuyqe@lenovo.com

Abstract

Scene text segmentation is a critical preprocessing step in various text-based applications. Specialist text segmentation methods, often relying on a detect-then-segment paradigm, tend to exhibit reduced robustness and can lead to cascading errors. The introduction of the Segment Anything Model (SAM) has revolutionized general segmentation by leveraging vision foundation models. However, SAM still falls short when applied to domain-specific tasks such as scene text segmentation. To bridge this gap between SAM and specialized scene text segmentation approaches, we propose **ST-SAM** (**S**cene **T**ext **SAM**), a parameter-efficient fine-tuning framework tailored to adapt SAM for high-quality scene text segmentation without relying on explicit text detection. ST-SAM incorporates a multimodal prompting mechanism: a lightweight visual encoder generates multi-scale spatial features to provide precise visual context; and textual prompts generated by a large language model offer high-level semantic guidance. We demonstrate the advantages of the proposed ST-SAM as follows: (1) ST-SAM achieves new state-of-the-art performance on multiple scene text segmentation benchmarks, including 85.30% fgIoU on Total-Text and 91.03% fgIoU on TextSeg, outperforming both specialist and generalist models. (2) ST-SAM enables effective domain adaptation by flexibly adapting the general SAM architecture to the domain of scene text. (3) By discarding the detect-then-segment pipeline, ST-SAM simplifies the inference process while still achieving robust performance on complex text cases.

Introduction

Scene text segmentation aims to segment text from complex scene images, by categorizing each pixel as either text foreground or no-text background, which plays a crucial role in various text-based applications (Azadi et al. 2018; Ulyanov, Vedaldi, and Lempitsky 2018; Wang et al. 2023b; Qu et al. 2023; Zhu et al. 2023; Zhang et al. 2025), thereby enhancing the versatility and utility of image processing in various tasks. Deep learning techniques have demonstrated considerable potential in text segmentation within scene images. Conventional approaches often follow a "*detect-then-segment*" paradigm, which relies heavily on accurately located text regions. As illustrated in Fig. 1, methods like TexRNet (Xu et al. 2021), PGTSNet (Xu et al. 2022), TFT (Yu et al. 2023a) and EAFormer (Yu et al. 2024) employ specialized text detection modules to enhance segmentation accuracy by specifically targeting text areas.

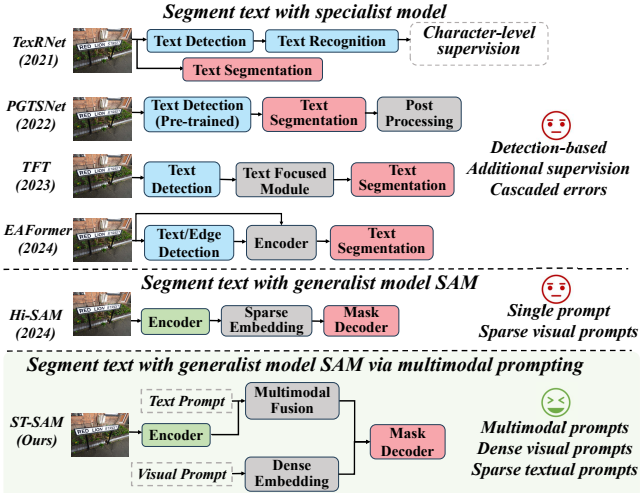

Figure 1: Conceptual comparison between our approach ST-SAM and prior text segmentation methods.

Despite recent progress, existing methods face two key limitations. First, these approaches heavily rely on the accuracy of the first-stage text detection. Since detection serves as the prerequisite for subsequent segmentation, any misalignment or error in locating text regions, particularly under complex conditions such as curved, dense, or stylized text, can lead to significant degradation in segmentation quality. This strong dependence limits the robustness and generalization capability of existing frameworks. Second, most frameworks struggle to adapt to diverse text layouts and appearance, limiting their practical applicability.

*Corresponding author

In contrast, generalist models such as the Segment Anything Model (SAM) (Kirillov et al. 2023) have demonstrated remarkable generalization capabilities, offering a promising new direction for overcoming the limitations of detection-dependent methods. SAM represents a paradigm shift in segmentation by leveraging prompt-based inference and vision foundation models, enabling flexible segmentation across a wide range of domains. However, despite its strong generalization ability, SAM underperforms in domain-specific tasks that require fine-grained perception, such as scene text segmentation where subtle structural cues and character-level details are crucial (Chen et al. 2023; Xie et al. 2024). Hi-SAM (Ye et al. 2024) addresses this gap by adapting SAM for text segmentation. They use embedded image features as sparse prompts, thereby improving the model's capability to segment text within images. Yet, its reliance on image-derived prompts alone limits its ability to resolve intricate text details, which are critical for handling irregular, disjointed text strokes and structures. This highlights a key challenge: designing an optimal architecture with a generalist model for high-quality text segmentation remains unresolved in this field.

In this paper, we propose **ST-SAM** to unleash the power of **SAM** for **S**cene **T**ext segmentation by *1) preserving the spatial integrity of text objects through multi-scale visual features and 2) enhancing the semantic understanding of textual content.* **To preserve the spatial integrity of text objects**, ST-SAM employs a lightweight semantic segmentation model, such as SegFormer (Xie et al. 2021), to generate multi-scale fused visual features. These features serve as dense visual prompts, supplying SAM with rich spatial context. By incorporating these dense visual prompts, ST-SAM significantly improves its capacity to segment text within scene images and resolves complex text details that are difficult to capture with traditional methods. **To enhance the semantic understanding of textual content**, we leverage the advanced language generation capabilities of Large Language Models (LLMs) first to generate general text-related prompts, and then complement them with scenario-specific prompts. This approach differs from relying on generic language prompts such as "*Text*" (Yu et al. 2023b) or predefined labels like "*Scene text*" (Zeng et al. 2024). The sparse textual prompt, in conjunction with the dense visual input, guides the segmentation process.

Additionally, we introduce a cross-modal alignment module to capture fine-grained, locality-sensitive image features that align with the semantic content of the textual prompts. This further enhances SAM's ability to accurately identify and segment text within scene images. As a result, ST-SAM demonstrates robustness in handling challenging cases without relying on a dedicated text detection module, achieving significant performance improvements over previous specialized models. Notably, ST-SAM improves the foreground Intersection over Union (fgIoU) by 2.57% on the Total-Text dataset and increases the F-score by 11.88% on a subset of particularly difficult cases. Furthermore, by eliminating the need for a separate text detection module, ST-SAM achieves an inference speed that is approximately 8.8 times faster than TexRNet (Xu et al. 2021).

In summary, the contributions of ST-SAM are as follows.

- We pioneer the integration of multi-scale fused **dense visual prompts** from a lightweight, specialized model. This design provides ST-SAM with comprehensive spatial information about text regions and edges, significantly enhancing its capacity to understand and segment text within scene images.
- To improve the semantic understanding of textual context, we propose **sparse textual prompts** that leverage LLM's language-generative capabilities to create semantically rich inputs. To align visual and textual features, we design a **cross-modal alignment module** that captures fine-grained image features with text prompts. This synergy sharpens ST-SAM's attention on textual content and improving segmentation accuracy.
- We conduct comprehensive experiments on five different datasets, demonstrating that ST-SAM achieves *state-of-the-art* results across all benchmarks.

Related Work

Scene Text Segmentation. Unlike general semantic and instance segmentation (Liu et al. 2024, 2018; Yu et al. 2018; Long, Shelhamer, and Darrell 2015; Strudel et al. 2021), scene text segmentation is typically formulated as a binary task. Since the ICDAR challenge (Karatzas et al. 2011), several datasets and methods have emerged (Ch'ng and Chan 2017; Bonechi et al. 2019, 2020). To address limited pixel-level labels, semi-supervised learning (Wang et al. 2021) and prior-guided models (Xu et al. 2021, 2022) leverage character or region priors to improve accuracy. Joint detection–segmentation frameworks (Yu et al. 2023a) and edge-aware transformers (Yu et al. 2024) further enhance boundary quality. However, most approaches still follow a detect–then–segment pipeline, which remains sensitive to complex backgrounds and diverse text structures.

Segment Anything Model and Applications. The Segment Anything Model (SAM) (Kirillov et al. 2023) is a category-agnostic, large-scale pre-trained segmentation model. HQ-SAM (Ke et al. 2023) improves its output quality by adding a learnable token and lightweight refinement layers. SAM has been adapted across domains, such as medical segmentation via LoRA-based or adapter-based variants like SAMed and MedSAM (Zhang and Liu 2023; Wu et al. 2023; Hu et al. 2021). In remote sensing, RSPrompt (Chen et al. 2024) designs anchor- and query-based prompts for SAM-driven instance segmentation. Motivated by these advancements, we explore applying SAM to scene text segmentation.

Prompt Learning. Recently, prompt learning with generalist models has attracted growing interest, forming a "pretraining–prompting" paradigm that reduces semantic gaps between pretraining and downstream tasks. In vision–language models, CoOp and CoCoOp (Zhou et al. 2022b,a) learn continuous prompts from few-shot data, while (Menon and Vondrick 2023) employs large language models to generate richer class descriptions. Beyond textual prompts, VPT (Jia et al. 2022) introduces visual prompts for

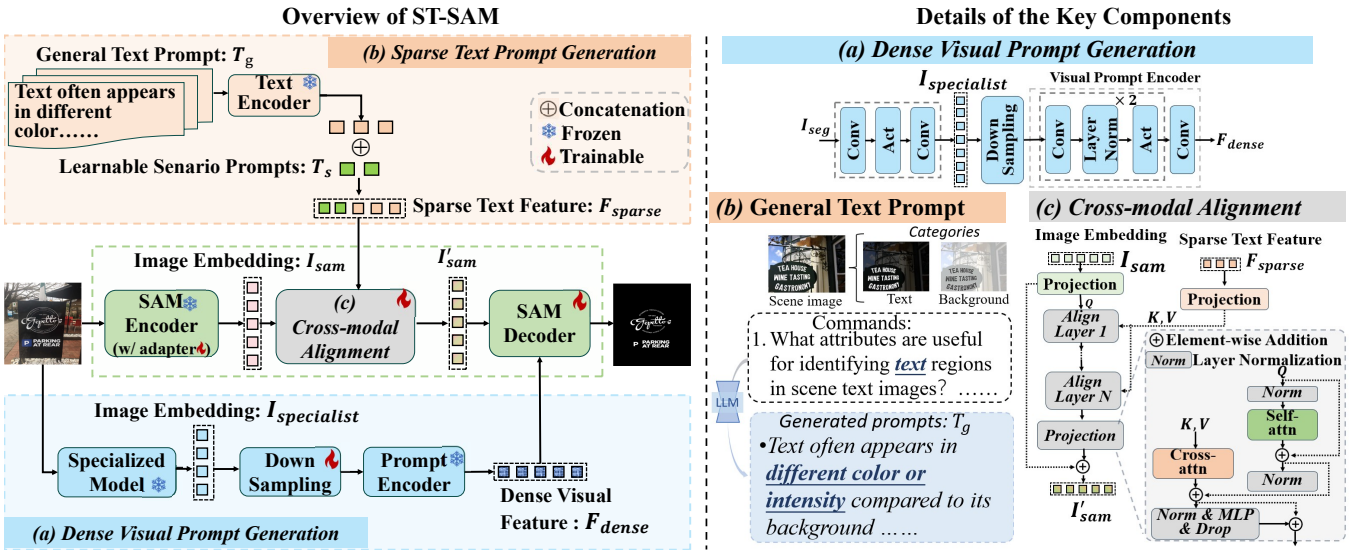

Figure 2: **Left:** ST-SAM overview based on SAM with trainable adapters. **Right:** Details of the key components: a) Dense visual prompt generation. b) Sparse text prompt generation. c) Cross-modal alignment module.

ViTs (Dosovitskiy et al. 2020), and LoGoPrompt (Shi and Yang 2023) shows that synthetic text images can act as effective visual prompts for improving classification.

Exploring SAM for Text Segmentation. Recently, Hi-SAM (Ye et al. 2024) transforms SAM into a hierarchical text segmentation model by incorporating a self-prompt module and a high-resolution mask decoder. The potential of using text, points, boxes, or masks as these initial prompts within Hi-SAM has yet to be thoroughly explored. In this paper, our approach involves creating text-specific prompts that are carefully designed to leverage SAM's inherent strengths and the unique characteristics of text. Unlike Hi-SAM, which uses image embeddings as sparse prompts, ST-SAM employs a multimodal prompting mechanism.

Methodology

Overview

Fig. 2 illustrates the structure of our ST-SAM model, which is designed to address scene text segmentation through a multimodal prompting approach. In the subsequent sections, we will delve into the specifics of how dense visual prompts are created, the generation of text prompts using LLMs, and the intricate process of integrating text prompts with the image embeddings to achieve optimal segmentation outcomes.

Dense Visual Prompt Generation

Our proposed approach capitalizes on the strengths of specialized semantic segmentation models to supply ST-SAM with multi-scale dense visual prompts. The core of our method lies in generating category-specific prompts that guide ST-SAM to focus on textual content within scene images. To accomplish this, we propose a *generalist-specialist* framework. Within this framework, SAM functions as a generalist for category-agnostic segmentation. Concurrently, a specialized semantic segmentation model, such as SegFormer, which has a relatively limited number of trainable parameters, acts as a specialist for scene text segmentation. SegFormer autonomously generates essential visual prompts for SAM, thereby significantly enhancing segmentation accuracy.

Formally, given a scene text image I as input, it is simultaneously processed by both SAM's image encoder and SegFormer's image encoder Ψ_{Seg}, producing outputs $I_{sam} \in \mathbb{R}^{64\times64\times256}$ and multi-scale feature maps denotes as $I_{seg} = [I_1^{c_1\times h_1\times w_1}, I_2^{c_2\times h_2\times w_2}, I_3^{c_3\times h_3\times w_3}, I_4^{c_4\times h_4\times w_4}]$, where w_i, h_i represents spatial resolutions at different scales and c_i denotes their respective channel dimensions. SegFormer has been pre-trained on the target dataset, and in our approach, its encoder parameters are frozen, while the decoder remains trainable. Each feature map within I_{seg} is first processed by a multi-layer perceptron (MLP), projected into a unified embedding space, then upsampled to match the highest spatial resolution among the scales, and concatenated along the channel dimension. The concatenated features undergo dimensionality reduction via two convolutional layers interleaved with a ReLU activation function, yielding the specialist feature representation $I_{specialist}$. Subsequently, these features are downsampled via a dedicated downsampling module $\Psi_{DownConv}$ to achieve a resolution of 256×256 using an adaptive pooling layer. These condensed and spatially-aligned features are then input into the visual prompt encoder $\Psi_{Visual\text{-}Prompt}$, producing the final dense visual prompt representation F_{dense}. Mathematically, this process is expressed as:

$$I_{sam} = \Psi_{SAM\text{-}Encoder}(I) \tag{1}$$

$$I'_{seg} = Concat(Up(MLP(I_i))), \quad i = 1, \ldots, 4 \tag{2}$$

$$I_{specialist} = Conv(ReLU(Conv(I'_{seg}))) \tag{3}$$

$$F_{dense} = \Psi_{Visual\text{-}Prompt}(\Psi_{DownConv}(I_{specialist})) \quad (4)$$

The visual prompt encoder $\Psi_{Visual\text{-}Prompt}$ comprises two convolutional layers with a kernel size of 2×2 and a stride of 2, followed by one additional convolutional layer with a kernel size of 1×1. After each of the first two convolutions, there is a layer normalization and a $GELU$ activation layer. This process produces F_{dense}, which has a spatial size of $64 \times 64 \times 256$.

Sparse Text Prompt Generation with LLMs

While dense visual prompts offer detailed spatial information crucial for preserving image features, they fall short in providing the semantic understanding necessary for accurate object recognition and segmentation, particularly in the complex task of scene text segmentation. To bridge this gap, we design a hybrid sparse text prompt mechanism that leverages both the semantic prior from LLM-generated prompts and scenario-adaptive learnable queries.

As illustrated in Fig. 2 (b), for the "*Text*" category, we first construct a set of general text prompts by querying LLMs (Achiam et al. 2023) with task-specific instructions, such as "*What attributes are useful for identifying text regions in scene text images?*". These generated sentences, denoted as T_g, capture general semantic attributes of text (e.g., edge sharpness, character alignment, or background contrast) without relying on any ground-truth annotations. To enhance adaptability across datasets and scenarios, we introduce L_1 learnable queries $Q_{learn} \in \mathbb{R}^{L_1 \times 512}$, which act as scenario-specific prompts. These queries are optimized jointly with the segmentation network, allowing the model to capture scenario-specific characteristics, such as typical font sizes, orientations, or background patterns.

The general text prompts T_g are first processed by a frozen CLIP text encoder $Text_{enc}(\cdot)$. For each input sentence T_i, the encoder produces a 1×512 feature vector representing its semantic meaning. With L_2 generated prompts, we obtain a semantic feature matrix $Text_{enc}(T_g) \in \mathbb{R}^{L_2 \times 512}$. We then concatenate the learnable queries and the encoded textual features along the sequence dimension:

$$F_{concat} = \left[\, Q_{learn} \,;\, Text_{enc}(T_g) \,\right] \in \mathbb{R}^{L \times 512} \quad (5)$$

where $[\,\cdot\,;\,\cdot\,]$ denotes feature concatenation, $L = L_1 + L_2$.

Finally, F_{concat} is projected to the target sparse representation through a trainable linear layer followed by a ReLU activation:

$$F_{sparse} = TextAffine(F_{concat}) \in \mathbb{R}^{L \times 256}, \quad (6)$$

This sparse representation encodes both the general semantic priors from LLM-generated prompts and the dataset-specific characteristics from learnable queries, effectively guiding ST-SAM to focus on textual regions under diverse scenarios.

Cross-modal Alignment

To effectively highlight the fine-grained information in response to the coarse text region for subsequent mask decoding, we design a cross-modal alignment module to dynamically integrate detailed semantic details from textual features into visual features. As shown in Fig. 2 (c), this is achieved through the cross-attention mechanism inherent in the Transformer (Vaswani et al. 2017) architecture, which effectively captures the interactions between image embeddings (Q) and text embeddings (K, V). The process involves calculating the aligned image feature $\hat{I} \in \mathbb{R}^{64 \times 64 \times 256}$, which is crucial for transferring semantic information from a general image context to the specific text-instance level. This is formulated as:

$$\hat{I} = \Psi_{CA}(Q = I_{sam}, K = F_{sparse}, V = F_{sparse}) \quad (7)$$

where Ψ_{CA} denotes the cross-modal alignment layer.

Utilizing the aligned visual features, the original image embedding I_{sam} is enhanced by incorporating $\hat{I}$ to generate text-aware localized embeddings $I'_{sam} \in \mathbb{R}^{64 \times 64 \times 256}$. The embeddings are then fed into the subsequent mask decoder $\Psi_{SAM\text{-}Decoder}$:

$$M = \Psi_{SAM\text{-}Decoder}(I_{sam} + \hat{I}, F_{dense}) \quad (8)$$

where $M \in \mathbb{R}^{256 \times 256 \times 1}$ denotes the predicted masks from mask decoder. To obtain high-resolution masks, M is upsampled using two transposed convolutional layers with a kernel size of 2×2 and a stride of 2, followed by four convolutional layers with a kernel size of 3×3, a stride of 1, and padding of 1, along with MLP layers. This process results in the final output $M' \in \mathbb{R}^{1024 \times 1024 \times 1}$. The cross-modal alignment module ensures that the mask decoder operates on an enriched representation conditioned on both visual and textual modalities, thereby enhancing the precision of the segmentation process.

Experiments

Experiments Setup

Datasets: We conduct experiments on the following datasets. **Total-Text** (Ch'ng and Chan 2017) comprises 1,555 images, with 1,255 designated for training and 300 for testing. The texts in this collection exhibit a wide range of orientations. **TextSeg** (Xu et al. 2021) is a large-scale fine-annotated text dataset. It comprises 4,024 images which are divided into 2,646 for training, 340 for validation, and 1,038 for testing. **ICDAR13 FST** (Karatzas et al. 2013) contains merely 462 images with 229 for training and 233 for testing. **HierText** (Long et al. 2022) comprises of 8,281 training images, 1,724 validation images, and 1,634 testing images. The dataset features dense and small text instances in scene and document images.

Evaluation Metric: To assess the proposed approach and ensure consistency with previous text segmentation methods, we utilize the foreground Intersection-over-Union (fgIoU) and the F-score of foreground pixels as the evaluation metric.

Implementation Details: In our experiments, we maintain an image size of 1024×1024, consistent with the original input size for the SAM model. We enhance training samples with data augmentation techniques including random rotation, color jittering, and large-scale jittering. For optimization, we use the AdamW optimizer with a learning rate of $1e-4$ and set the total number of training

	Method	Total-Text		TextSeg		ICDAR13_FST		Inference
		fgIOU	F-score	fgIOU	F-score	fgIOU	F-score	Time (s/img)
Specialist models	HRNetV2-W48 + OCR (Yuan et al. 2020)	76.23	83.20	85.98	91.80	72.45	83.00	-
	TexRNet + HRNetV2-W48 (Xu et al. 2021)	78.47	84.80	86.84	92.40	73.38	85.00	1.32
	SegFormer-B5 †(Xie et al. 2021)	79.96	88.87	87.57	93.37	65.31	79.02	0.22
	PGTSNet (Xu et al. 2022)	79.10	84.70	-	-	-	-	-
	TFT (Yu et al. 2023a)	82.10	90.20	87.11	93.10	72.71	84.50	-
	TextFormer (Wang et al. 2023a)	81.56	88.70	87.42	93.30	72.27	83.80	0.42*
	EAFormer (Yu et al. 2024)	82.73	90.60	88.06	93.90	72.63	84.00	0.47*
Generalist models	SAM-B †(Kirillov et al. 2023)	75.06	85.75	86.82	92.94	65.70	79.30	0.10
	SAM-L †(Kirillov et al. 2023)	78.46	87.93	87.95	93.59	68.03	80.97	0.24
	SAM-H †(Kirillov et al. 2023)	81.16	89.61	88.39	93.84	73.54	84.75	0.42
	UPOCR ‡(Peng et al. 2024)	81.24	88.50	88.76	94.04	73.68	84.92	0.16
	Hi-SAM-B (Ye et al. 2024)	80.93	86.25	87.15	92.81	-	-	0.11
	Hi-SAM-L (Ye et al. 2024)	84.59	88.69	88.77	93.79	-	-	0.26
	Hi-SAM-H (Ye et al. 2024)	84.86	89.68	88.96	93.87	-	-	0.45
	SAM2-B+ †(Ravi et al. 2024)	61.25	75.97	78.18	87.75	59.43	74.55	0.05
	SAM2-L †(Ravi et al. 2024)	67.87	80.86	81.87	90.03	69.80	82.21	0.05
	ST-SAM-B (Ours)	82.36	90.33	88.81	94.07	72.62	84.14	0.15
	ST-SAM-L (Ours)	<u>85.14</u>	<u>91.97</u>	<u>90.67</u>	<u>95.11</u>	<u>74.87</u>	<u>85.63</u>	0.40
	ST-SAM-H (Ours)	**85.30**	**92.07**	**91.03**	**95.65**	**75.17**	**85.92**	0.59

Table 1: Performance comparison with existing methods. The results marked with † are fine-tune by us. The results marked with ‡ are reproduced by us. **Bold** and <u>underline</u> represent the best and second-best performance, respectively. The results of $*$ are from (Yu et al. 2024).

epochs to 70 across all datasets, with a batch size of 4. The cross-modal alignment module is composed of three transformer layers, each with four heads. We conducted three sets of experiments using the image encoders ViT-B, ViT-L, and ViT-H, referred to as ST-SAM-B, ST-SAM-L, and ST-SAM-H, respectively. All our experiments are carried out on NVIDIA A800 GPUs. The number of learnable scenario-specific prompts is set to 1 by default. *See the supplementary material for more details.*

Training of ST-SAM: In our experiments, we use the SegFormer MiT-B1 lightweight encoder for ST-SAM-B and the SegFormer MiT-B5 for ST-SAM-L / ST-SAM-H. *See the supplementary material for a detailed explanation of the selection of SegFormer models.* During the training process, the learnable modules of ST-SAM include: 1) the adapters in the ViT encoder, 2) the MLP layer of the specialized model, 3) the cross-modal alignment module, 4) the mask decoder, and 5) the affine layer of the text encoder and the learnable scenario-specific prompts. Following the experimental setup of early method (Ye et al. 2024), the loss function is defined as a linear combination of Focal loss, Dice loss, and MSE loss, with a ratio of 20:1:1.

Method	fgIoU	F-score
TexRNet + HRNetV2-W48 (Xu et al. 2021)	55.50	65.64
Hi-SAM-B (Ye et al. 2024)	73.39	81.34
ST-SAM-B (Ours)	**75.69**	**86.17**

Table 2: Performance comparison on HierText.

Performance Comparison

The performance comparison results between our method and the existing representative methods are shown in Tab. 1. It is evident that specialized methods, which often depend on well-designed text detection modules, did not achieve the best results across the datasets. In contrast, ST-SAM, which leverages the powerful segmentation capabilities of SAM, has demonstrated a marked superiority over all alternative methods. This success can largely be attributed to the generalist models' ability to provide a rich feature set that transcends the constraints of dataset size, enabling more accurate and reliable text segmentation. **On the three Latin datasets**, our ST-SAM-H method showed average improvements of 2.69%/1.71% in fgIoU/F-score over the SOTA specialist model EAFormer (Yu et al. 2024). We also show results on small and thin text. Table 2 illustrates the performance **on HierText**, showcasing superior outcomes with a 2.3%/4.83% improvement in fgIoU/F-score compared to Hi-SAM. This improvement is attributed to the rich spatial and semantic information provided by our multimodal prompting technique.

We additionally fine-tune SAM (Kirillov et al. 2023) and the recently introduced SAM2 (Ravi et al. 2024) on the target training datasets. For a fair comparison, we follow the same fine-tune strategy as ST-SAM. SAM demonstrates inferior performance compared to our method, while SAM2, designed particularly for video segmentation, exhibits even poorer performance in this specific task. We attribute this underperformance to the factor that SAM2 replaces the Vision Transformer (ViT) (Dosovitskiy et al. 2020) with Hiera (Ryali et al. 2023) to achieve faster processing times.

Visualization Analysis. As illustrated in Fig. 3, we com-

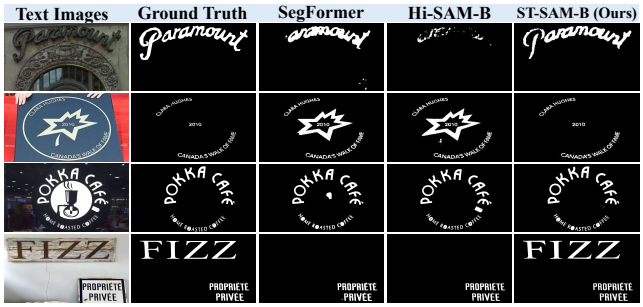

Figure 3: Qualitative comparisons on Total-Text and TextSeg dataset. Zoom in for a better view.

pare the performance of our ST-SAM model with state-of-the-art methods such as SegFormer and Hi-SAM on the Total-Text and TextSeg datasets. ST-SAM produces accurate mask predictions with reduced noise. Notably, in the first and fourth rows, where other methods fail to detect certain characters, ST-SAM successfully identifies all of them. In the second and third rows, while other methods are prone to confusing text-like areas, ST-SAM maintains high precision due to the detailed spatial and semantic information provided by the integration of dense visual and sparse text prompts. *More visualization results are shown in the supplementary material.*

Performance on Difficult Scenarios

To further demonstrate the effectiveness of ST-SAM in difficult scenarios, we compare it against previous state-of-the-art frameworks on a selected subset of the Total-Text dataset, which mainly consists of multi-oriented and densely packed text. Specifically, we selected 185 difficult samples from the Total-Text dataset. As shown in Tab. 3, compared with specialist models, ST-SAM outperforms the other frameworks by a significant margin. This result underscores the robustness of our framework, especially in handling difficult cases. *Visualization comparisons are shown in supplementary.*

Method	fgIoU	F-score
TexRNet + DeeplabV3+ (Xu et al. 2021)	71.55	77.05
TexRNet + HRNetV2-W48 (Xu et al. 2021)	74.54	77.42
UPOCR (Peng et al. 2024)	76.34	86.58
ST-SAM-B (Ours)	**80.67**	**89.30**

Table 3: Comparisons with previous specialist and generalist frameworks on **difficult scenarios**.

Generalization Ability

We conduct experiments on the task of real-to-real adaptation, specifically comparing our approach with specialist and generalist methods, as shown in Tab. 4. The models are trained solely on TextSeg and evaluated on TotalText. It is noteworthy that TextSeg mainly comprises artificially designed text, whereas TotalText encompasses natural scene text. Taking advantage of the strong generalization capabilities of SAM and scenario-adaptive prompts, our method outperforms specialist models with an average gap of 5.8% and a gap of 4.29% compared to Hi-SAM, underscoring its efficacy in domain adaptation. Despite the differences in text complexity and contextual diversity between the training and evaluation datasets, ST-SAM demonstrates its generalization ability in various real-world scenarios.

Method	TextSeg → Total-Text
Segformer-B1 (Xie et al. 2021)	80.78
TexRNet + DeeplabV3+ (Xu et al. 2021)	80.54
TexRNet + HRNetV2-W48 (Xu et al. 2021)	80.96
Hi-SAM-B (Ye et al. 2024)	82.27
ST-SAM-B (Ours)	**86.56**

Table 4: Comparisons with previous specialist and generalist frameworks on real-to-real adaptation. F-score (%) is reported.

Ablation Studies

To assess the performance of each component in ST-SAM, we perform ablation experiments on the ICDAR13_FST dataset, as shown in Tab. 5.

Impact of the Adapter. In the row of 1-2 of Tab. 5, we analysis the impact of adopting an adapter tuning approach. By incorporating two trainable adapter modules into each ViT block, we introduce a few trainable parameters into the encoder. This modification leads to significantly improved performance compared to the baseline model.

Impact of Visual Prompting. In rows 2-3 and 5-8 of Tab. 5, we illustrate the impact of dense visual prompting within ST-SAM. By generating dense visual prompts that correspond to the image's embedding spatial dimensions, our method achieves improvements of 1.34% (row 2 vs 3) and 1.70% (row 5 vs 8) of fgIoU, respectively. These results indicate that dense visual prompts are essential for incorporating detailed spatial information into the generalist model.

Impact of Text Prompting with LLM. We illustrate the impact of the design of the text prompting within ST-SAM in rows 4-5 and 7-8 of Tab. 5. The label "w/o LLM" indicates that the "*Text*" input is encoded solely using the text encoder, while "w/ LLM" signifies that text-related prompts are generated using GPT-4 before being input into the text encoder. Generating more semantically rich prompts provides valuable guidance to ST-SAM for capturing finer details. This results in an overall improvement (row 7 vs 8) of 1.08% in fgIoU. We also evaluate the capability of other open-source LLMs, such as Qwen-2.5 (Qwen et al. 2025) *in the supplementary material*. Our results show that performance improvements stem from (1) sparse text prompt generation strategy and (2) their ability to capture fine-grained image details—not the choice of specific LLMs.

Impact of Cross-modal Alignment Module. The results indicate that simple concatenation reduces the model's performance (row 3 vs 6). In contrast, using the cross-modal alignment improves the model's performance (row 6 vs 8). In Fig. 4, we present text-aware localized embeddings I'_{sam}

#	Adapter	Visual Prompting	Text Prompting	Cross-modal Alignment	Scenario-specific prompt	ICDAR13_FST fgIoU	ICDAR13_FST F-score
1	✗	✗	✗	✗	✗	65.70	79.30
2	✓	✗	✗	✗	✗	68.18	81.08
3	✓	✓	✗	✗	✗	69.52	82.02
4	✓	✗	✓, w/o LLM	✓	✗	68.68	81.43
5	✓	✗	✓, w/ LLM	✓	✗	69.48	81.99
6	✓	✓	✓, w/ LLM	✗, concat	✗	68.49	81.30
7	✓	✓	✓, w/o LLM	✓	✗	70.10	82.42
8	✓	✓	✓, w/ LLM	✓	✗	71.18	83.17
9	✓	✓	✗	✓	✓	70.99	83.03
10	✓	✓	✓, w/ LLM	✓	✓	**72.62**	**84.14**

Table 5: Ablation experiments on ICDAR13_FST. The term "concat" refers to the concatenation of text prompts with the sparse prompts of SAM. Note that Cross-modal Alignment is conducted between the image embeddings and sparse text features.

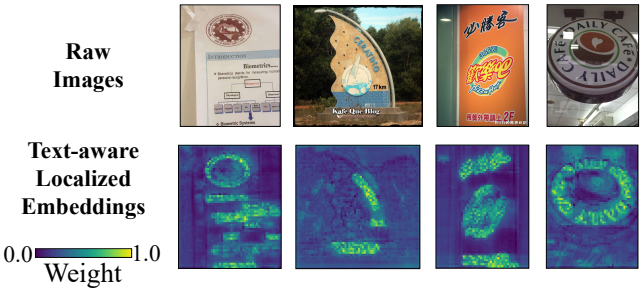

Figure 4: Visualization of text-aware localized embeddings I'_{sam} after cross-modal alignment module.

after cross-modal alignment module. The visualization reveals that the features within I'_{sam} cover precise text location details, whether the text appears in cluttered scenes or spans various orientations. This demonstrates the versatility of ST-SAM in identifying text regions and provides crucial insights for the subsequent mask decoding phase.

Impact of Scenario-specific Prompts. Rows 8-10 evaluate different strategies for sparse prompts. Row 8 uses only LLM-generated prompts (general semantics), row 9 uses only learnable scenario-specific prompts, and row 10 combines both. Results show that both types of prompts are beneficial: general prompts contribute semantic prior knowledge, while learnable scenario-specific prompts adapt to instance-specific features. Their combination yields the best performance (72.62% fgIoU, 84.14 F-score), underscoring the complementarity of the two prompting strategies.

Analysis of ST-SAM's Inference Speed and Trainable Parameters. We present the inference time in Tab. 1 and the number of trainable parameters in Fig. 5. We evaluate the inference time of all available methods on one single A800 GPU card. ST-SAM demonstrates an inference speed comparable to other generalist models except SAM2. Regarding trainable parameters, our method uses a parameter-efficient fine-tuning approach, enabling ST-SAM to achieve superior performance with fewer parameters.

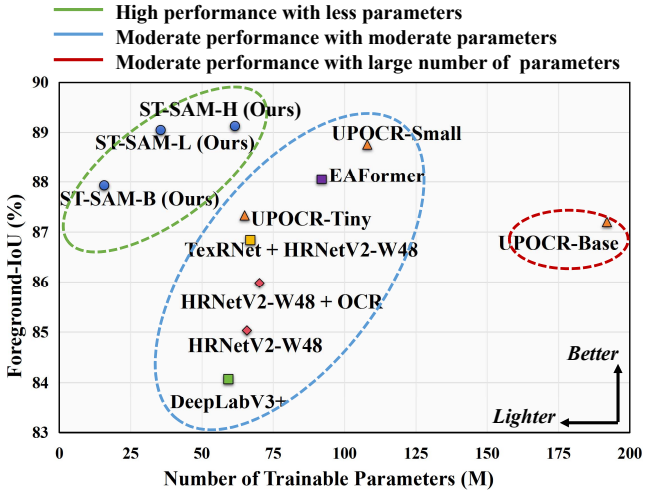

Figure 5: Comparison of previous methods in the number of parameters and text segmentation performance in fgIoU.

Conclusion

In this paper, we introduce ST-SAM, a novel multimodal prompting approach for scene text segmentation that builds on the SAM model. Our approach leverages a lightweight specialized model to generate multi-scale fused visual prompts, providing comprehensive spatial information. Additionally, we harness LLM's advanced generation capabilities to create general prompts, then complement them with scenario-specific prompts, enriching the semantic input for ST-SAM. The integration of dense visual prompts and sparse text prompts, along with the cross-modal alignment module, enhances the segmentation accuracy of ST-SAM. Through extensive experiments, we have demonstrated the superiority of ST-SAM over other state-of-the-art specialist and generalist models. Furthermore, ST-SAM exhibits robustness in handling difficult cases and effectiveness in domain adaptation. In future work, we will extend this framework to a broader range of pixel-level OCR tasks.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Azadi, S.; Fisher, M.; Kim, V. G.; Wang, Z.; Shechtman, E.; and Darrell, T. 2018. Multi-content gan for few-shot font style transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7564–7573.

Bonechi, S.; Andreini, P.; Bianchini, M.; and Scarselli, F. 2019. COCO_TS dataset: pixel–level annotations based on weak supervision for scene text segmentation. In *International Conference on Artificial Neural Networks*, 238–250. Springer.

Bonechi, S.; Bianchini, M.; Scarselli, F.; and Andreini, P. 2020. Weak supervision for generating pixel–level annotations in scene text segmentation. *Pattern Recognition Letters*, 138: 1–7.

Chen, K.; Liu, C.; Chen, H.; Zhang, H.; Li, W.; Zou, Z.; and Shi, Z. 2024. RSPrompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model. *IEEE Transactions on Geoscience and Remote Sensing*.

Chen, T.; Zhu, L.; Deng, C.; Cao, R.; Wang, Y.; Zhang, S.; Li, Z.; Sun, L.; Zang, Y.; and Mao, P. 2023. Sam-adapter: Adapting segment anything in underperformed scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3367–3375.

Ch'ng, C. K.; and Chan, C. S. 2017. Total-text: A comprehensive dataset for scene text detection and recognition. In *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*, volume 1, 935–942. IEEE.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *ICLR*.

Jia, M.; Tang, L.; Chen, B.-C.; Cardie, C.; Belongie, S.; Hariharan, B.; and Lim, S.-N. 2022. Visual prompt tuning. In *European Conference on Computer Vision*, 709–727. Springer.

Karatzas, D.; Mestre, S. R.; Mas, J.; Nourbakhsh, F.; and Roy, P. P. 2011. ICDAR 2011 robust reading competition-challenge 1: reading text in born-digital images (web and email). In *2011 international conference on document analysis and recognition*, 1485–1490. IEEE.

Karatzas, D.; Shafait, F.; Uchida, S.; Iwamura, M.; i Bigorda, L. G.; Mestre, S. R.; Mas, J.; Mota, D. F.; Almazan, J. A.; and De Las Heras, L. P. 2013. ICDAR 2013 robust reading competition. In *2013 12th international conference on document analysis and recognition*, 1484–1493. IEEE.

Ke, L.; Ye, M.; Danelljan, M.; Liu, Y.; Tai, Y.-W.; Tang, C.-K.; and Yu, F. 2023. Segment anything in high quality. In *NeurIPS*.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4026.

Liu, Q.; Cho, J.; Bansal, M.; and Niethammer, M. 2024. Rethinking Interactive Image Segmentation with Low Latency High Quality and Diverse Prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3773–3782.

Liu, S.; Qi, L.; Qin, H.; Shi, J.; and Jia, J. 2018. Path aggregation network for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8759–8768.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3431–3440.

Long, S.; Qin, S.; Panteleev, D.; Bissacco, A.; Fujii, Y.; and Raptis, M. 2022. Towards End-to-End Unified Scene Text Detection and Layout Analysis.

Menon, S.; and Vondrick, C. 2023. Visual classification via description from large language models. *ICLR*.

Peng, D.; Yang, Z.; Zhang, J.; Liu, C.; Shi, Y.; Ding, K.; Guo, F.; and Jin, L. 2024. UPOCR: Towards unified pixel-level ocr interface. In *Forty-first International Conference on Machine Learning*.

Qu, Y.; Tan, Q.; Xie, H.; Xu, J.; Wang, Y.; and Zhang, Y. 2023. Exploring stroke-level modifications for scene text editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 2119–2127.

Qwen; :; Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; Lin, H.; Yang, J.; Tu, J.; Zhang, J.; Yang, J.; Yang, J.; Zhou, J.; Lin, J.; Dang, K.; Lu, K.; Bao, K.; Yang, K.; Yu, L.; Li, M.; Xue, M.; Zhang, P.; Zhu, Q.; Men, R.; Lin, R.; Li, T.; Tang, T.; Xia, T.; Ren, X.; Ren, X.; Fan, Y.; Su, Y.; Zhang, Y.; Wan, Y.; Liu, Y.; Cui, Z.; Zhang, Z.; and Qiu, Z. 2025. Qwen2.5 Technical Report. arXiv:2412.15115.

Ravi, N.; Gabeur, V.; Hu, Y.-T.; Hu, R.; Ryali, C.; Ma, T.; Khedr, H.; Rädle, R.; Rolland, C.; Gustafson, L.; Mintun, E.; Pan, J.; Alwala, K. V.; Carion, N.; Wu, C.-Y.; Girshick, R.; Dollár, P.; and Feichtenhofer, C. 2024. SAM 2: Segment Anything in Images and Videos. *arXiv preprint arXiv:2408.00714*.

Ryali, C.; Hu, Y.-T.; Bolya, D.; Wei, C.; Fan, H.; Huang, P.-Y.; Aggarwal, V.; Chowdhury, A.; Poursaeed, O.; Hoffman, J.; et al. 2023. Hiera: A hierarchical vision transformer without the bells-and-whistles. In *International Conference on Machine Learning*, 29441–29454. PMLR.

Shi, C.; and Yang, S. 2023. Logoprompt: Synthetic text images can be good visual prompts for vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2932–2941.

Strudel, R.; Garcia, R.; Laptev, I.; and Schmid, C. 2021. Segmenter: Transformer for semantic segmentation. In *Pro-*

ceedings of the IEEE/CVF international conference on computer vision, 7262–7272.

Ulyanov, D.; Vedaldi, A.; and Lempitsky, V. 2018. Deep image prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 9446–9454.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, C.; Zhao, S.; Zhu, L.; Luo, K.; Guo, Y.; Wang, J.; and Liu, S. 2021. Semi-supervised pixel-level scene text segmentation by mutually guided network. *IEEE Transactions on Image Processing*, 30: 8212–8221.

Wang, X.; Wu, C.; Yu, H.; Li, B.; and Xue, X. 2023a. Textformer: component-aware text segmentation with transformer. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, 1877–1882. IEEE.

Wang, Y.; Xie, H.; Wang, Z.; Qu, Y.; and Zhang, Y. 2023b. What is the real need for scene text removal? exploring the background integrity and erasure exhaustivity properties. *IEEE Transactions on Image Processing*.

Wu, J.; Fu, R.; Fang, H.; Liu, Y.; Wang, Z.; Xu, Y.; Jin, Y.; and Arbel, T. 2023. Medical sam adapter: Adapting segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.12620*.

Xie, E.; Lyu, J.; Wu, D.; Shen, H.; and Zhou, Y. 2024. Char-SAM: Turning Segment Anything Model into Scene Text Segmentation Annotator with Character-level Visual Prompts. arXiv:2412.19917.

Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; and Luo, P. 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34: 12077–12090.

Xu, X.; Qi, Z.; Ma, J.; Zhang, H.; Shan, Y.; and Qie, X. 2022. BTS: a bi-lingual benchmark for text segmentation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 19152–19162.

Xu, X.; Zhang, Z.; Wang, Z.; Price, B.; Wang, Z.; and Shi, H. 2021. Rethinking text segmentation: A novel dataset and a text-specific refinement approach. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12045–12055.

Ye, M.; Zhang, J.; Liu, J.; Liu, C.; Yin, B.; Liu, C.; Du, B.; and Tao, D. 2024. Hi-SAM: Marrying Segment Anything Model for Hierarchical Text Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–16.

Yu, C.; Wang, J.; Peng, C.; Gao, C.; Yu, G.; and Sang, N. 2018. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, 325–341.

Yu, H.; Fu, T.; Li, B.; and Xue, X. 2024. EAFormer: Scene Text Segmentation with Edge-Aware Transformers. In *Proceedings of the European conference on computer vision (ECCV)*.

Yu, H.; Wang, X.; Niu, K.; Li, B.; and Xue, X. 2023a. Scene text segmentation with text-focused transformers. In *Proceedings of the 31st ACM International Conference on Multimedia*, 2898–2907.

Yu, W.; Liu, Y.; Hua, W.; Jiang, D.; Ren, B.; and Bai, X. 2023b. Turning a clip model into a scene text detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6978–6988.

Yuan, Y.; et al. 2020. Object-contextual representations for semantic segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, 173–190. Springer.

Zeng, G.; Zhang, Y.; Wei, J.; Yang, D.; Zhang, P.; Gao, Y.; Qin, X.; and Zhou, Y. 2024. Focus, Distinguish, and Prompt: Unleashing CLIP for Efficient and Flexible Scene Text Retrieval. In *Proceedings of the 32nd ACM International Conference on Multimedia*.

Zhang, K.; and Liu, D. 2023. Customized segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.13785*.

Zhang, Y.; Liu, C.; Wei, J.; Yang, X.; Zhou, Y.; Ma, C.; and Ji, X. 2025. Linguistics-aware Masked Image Modeling for Self-supervised Scene Text Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022a. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16816–16825.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022b. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337–2348.

Zhu, S.; Zhao, Z.; Fang, P.; and Xue, H. 2023. Improving scene text image super-resolution via dual prior modulation network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 3843–3851.