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

AG-VAS: Anchor-Guided Zero-Shot Visual Anomaly Segmentation with Large Multimodal Models

Zhen Qu[1, 2] Xian Tao[1,2,3,4†] Xiaoyi Bao[1,2] Dingrong Wang[1, 2]
ShiChen Qu[1, 2] Zhengtao Zhang[1,2,3] Xingang Wang[1, 2]
[1]Institute of Automation, Chinese Academy of Sciences
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences
[3]Casivision [4] Weiqiao-UCAS Science and Technology Park

Abstract

Large multimodal models (LMMs) exhibit strong task generalization capabilities, offering new opportunities for zero-shot visual anomaly segmentation (ZSAS). However, existing LMM-based segmentation approaches still face fundamental limitations: anomaly concepts are inherently abstract and context-dependent, lacking stable visual prototypes, and the weak alignment between high-level semantic embeddings and pixel-level spatial features hinders precise anomaly localization. To address these challenges, we present AG-VAS (Anchor-Guided Visual Anomaly Segmentation), a new framework that expands the LMM vocabulary with three learnable semantic anchor tokens—[SEG], [NOR], and [ANO], establishing a unified anchor-guided segmentation paradigm. Specifically, [SEG] serves as an absolute semantic anchor that translates abstract anomaly semantics into explicit, spatially grounded visual entities (e.g., holes or scratches), while [NOR] and [ANO] act as relative anchors that model the contextual contrast between normal and abnormal patterns across categories. To further enhance cross-modal alignment, we introduce a Semantic-Pixel Alignment Module (SPAM) that aligns language-level semantic embeddings with high-resolution visual features, along with an Anchor-Guided Mask Decoder (AGMD) that performs anchor-conditioned mask prediction for precise anomaly localization. In addition, we curate Anomaly-Instruct20K, a large-scale instruction dataset that organizes anomaly knowledge into structured descriptions of appearance, shape, and spatial attributes, facilitating effective learning and integration of the proposed semantic anchors. Extensive experiments on six industrial and medical benchmarks demonstrate that AG-VAS achieves consistent state-of-the-art performance in the zero-shot setting.

†Corresponding Author: taoxian2013@ia.ac.cn.
‡Project Page: https://github.com/xiaozhen228/AG-VAS

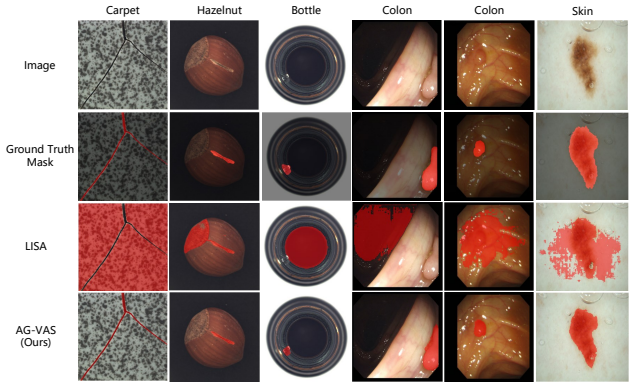


Figure 1. Comparison of zero-shot anomaly segmentation capabilities between LISA [21] and our AG-VAS.

1. Introduction

Zero-shot visual anomaly segmentation (ZSAS) aims to directly localize anomalous regions on the surfaces of unseen-category objects without the need for retraining. This capability is particularly important in data-scarce or privacy-sensitive scenarios, such as industrial defect localization [33, 34, 43] and medical image analysis [2, 36, 37].

Most existing ZSAS models [10–12, 19, 35, 36, 56] are built upon CLIP [38] due to its strong vision–language alignment capabilities. These models typically construct pairs of textual prompts representing normal and abnormal states, and align them with pixel-level patch embeddings to localize anomalies. Representative methods, such as AnomalyCLIP [56] and Bayes-PFL [36], further enhance the fine-grained anomaly perception via prompt learning strategies. However, constrained by the limited representation and understanding capabilities of CLIP, the ZSAS performance of these approaches appears to have reached a bottleneck. Moreover, they do not naturally produce binary segmentation masks, often requiring heuristic thresh-

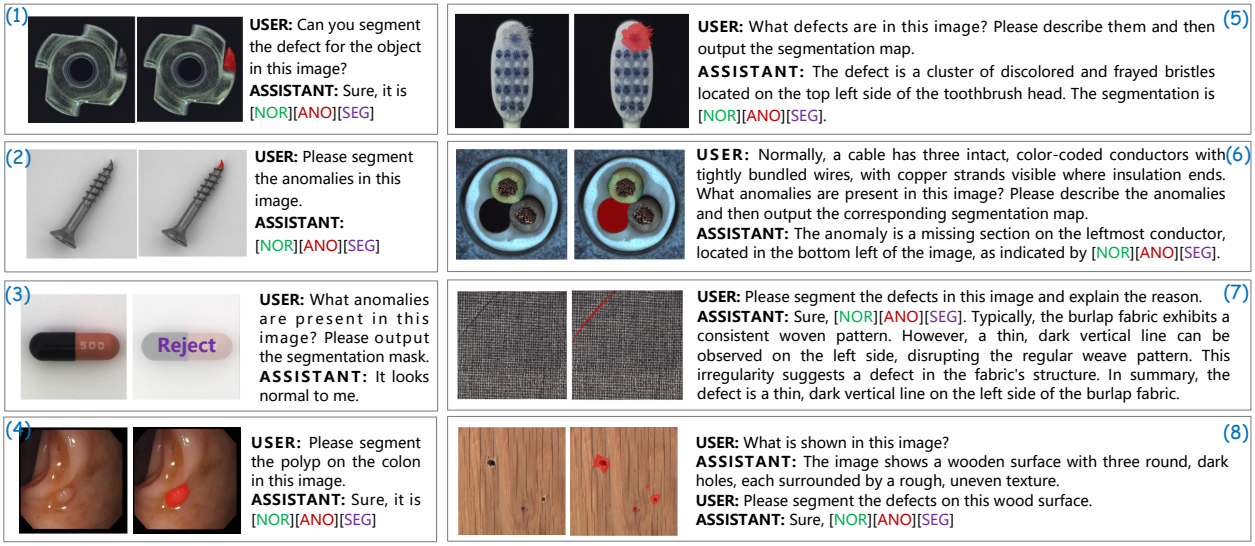


Figure 2. AG-VAS is an effective and efficient LMM for zero-shot anomaly segmentation and understanding. (Left: raw image; Right: segmentation result.) The visualization results are shown across the following scenarios: (1)–(4) direct segmentation, (5)–(6) describe-then-segment, (7) segment-then-explain, and (8) segmentation within dialogue.

olding or empirical tuning, which hinders practical deployment. These limitations motivate exploring large multimodal models (LMMs) with stronger reasoning capacity and richer world knowledge [13, 28, 39, 47, 57].

Recent anomaly detection studies [17, 18, 24, 49] leveraging LMMs have demonstrated remarkable performance in high-level visual understanding and conversational interaction. However, most existing works (e.g. Anomaly-OV [49]) focus on enhancing image-level textual descriptions or anomaly classification accuracy, while precise pixel-level responses, such as binary segmentation masks, remain largely unexplored. In this work, we aim to develop an end-to-end ZSAS framework built upon the pretrained LMMs, capable of understanding complex segmentation instructions and directly outputting binary masks.

LISA [21] is the first general-purpose image segmentation work to employ an *embedding-as-mask* scheme, extending visual comprehension to pixel-wise segmentation by integrating LMMs with dedicated segmenters (e.g. SAM [20]). However, when applied to industrial or medical ZSAS tasks, it often fails to correctly identify anomalous regions, sometimes even confusing foreground and background. For instance, given the instruction "`Please segment the anomalies in this image`", LISA may produce masks that mislocalize the abnormal regions, as illustrated in Figure 1. We attribute these failures in ZSAS to two main factors: 1) anomalies or defects are highly abstract concepts without concrete semantic references—for example, unlike a well-defined object such as "apple", anomalies can include "holes" or "scratches", making them difficult to map from textual descriptions; 2) the segmenter's pixel-level features are not well aligned with the LMM's visual-textual embedding space, limiting effective anomaly localization.

To address the above issues, we propose **AG-VAS**, an end-to-end ZSAS framework equipped with a set of learnable anchor tokens that encode contextual and world knowledge about anomalies. These anchors serve as semantic bridges between the LMM and the segmenter, enabling effective instruction-driven anomaly segmentation. Specifically, we extend the LMM vocabulary with special anchor tokens: the absolute semantic anchor [SEG] provides explicit cues regarding anomaly appearance, shape, and location, while the relative semantic anchors [NOR] and [ANO] capture contextual contrasts between normal and abnormal regions across diverse categories. To enhance cross-modal correspondence, we further introduce a *Semantic–Pixel Alignment Module* (SPAM), which aligns high-level semantic embeddings from the LMM with fine-grained pixel features from the segmenter. Finally, an *Anchor-Guided Mask Decoder* (AGMD) leverages these aligned representations for anchor-consistent mask generation, achieving binarized segmentation results.

To further enrich the model's understanding of anomaly-related semantics, we construct **Anomaly-Instruct20K**, an instruction-tuning dataset specifically designed for segmentation tasks. Unlike conventional instruction datasets focused on dialogue or visual question answering, Anomaly-Instruct20K injects structured world knowledge of anomalies into the LMM through diverse descriptions of defect appearance, shape, and spatial context. This enables the anchor tokens to learn precise and discriminative seman-

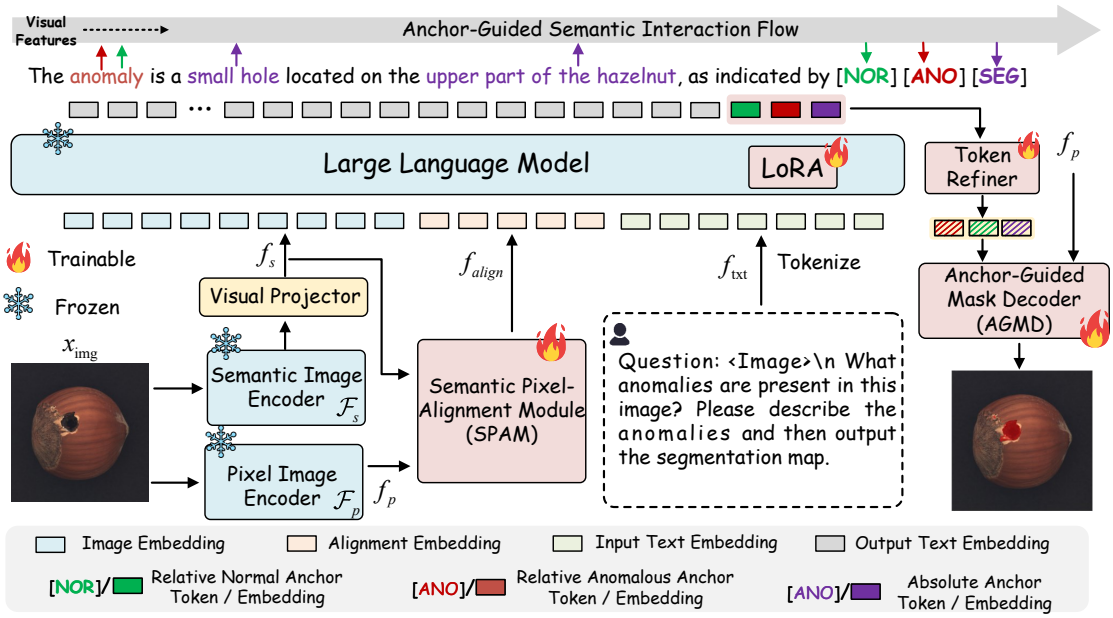


Figure 3. The pipeline of AG-VAS. We design two types of semantic anchors: the absolute anchor [SEG], which injects world and contextual knowledge about the appearance, structure, and location of anomalies (e.g., holes, scratches) into the segmentation process, and the relative normal/anomalous anchors [NOR]/[ANO], which guide the model to understand anomalies in relation to normal patterns. All parameters except those for the semantic/pixel image encoder, visual projector and LLM are trainable.

tic representations. Moreover, its instruction-based design allows AG-VAS to develop interactive segmentation behaviors and flexibly respond to various user prompts. Figure 2 demonstrates that AG-VAS produces high-quality masks under different segmentation instructions. In summary, our contributions are:

- We propose **AG-VAS**, an anchor-guided framework for ZSAS that directly produces binary masks through instruction-driven segmentation. It introduces absolute and relative semantic anchors to bridge the LMM and segmenter, and incorporates two modules: SPAM for aligning LMM and segmenter features, and AGMD for generating accurate, anchor-consistent masks.
- We construct Anomaly-Instruct20K, a segmentation-oriented instruction-tuning dataset that injects structured world knowledge of anomalies into the LMM, enhancing its understanding of defect-related semantics.
- We achieve state-of-the-art performance on multiple industrial and medical anomaly detection benchmarks, demonstrating the strong generalization ability and practical applicability of AG-VAS.

2. Related Work

2.1. Large Multimodal Models

Recent advances in LMMs can be broadly categorized into two paradigms: comprehension-oriented and perception-oriented approaches. Comprehension-oriented methods focus on enhancing visual understanding by aligning visual encoders with language models and improving representation quality through progressive training or multi-scale feature design [1, 15, 16, 23, 25, 28, 46]. Representative models such as BLIP-2 [23], Flamingo [1], LLaVA [28], and Qwen-VL [46] have substantially advanced cross-modal reasoning and instruction-following capabilities. More recently, techniques like chain-of-thought reasoning [29, 32] and mixture-of-experts architectures [27, 50] have further improved LMMs' ability to handle complex visual comprehension tasks. In contrast, perception-oriented LMMs emphasize fine-grained visual grounding, aiming to adapt language models for pixel-level localization tasks [6, 21, 26, 39, 41, 48, 54]. A promising direction is the embedding-as-mask paradigm, first introduced by LISA [21] and later extended by PixelLM [39] and PaDT [41]. While these methods achieve strong performance in general object segmentation, they remain suboptimal when applied to the ZSAS task.

2.2. Zero-Shot Anomaly Segmentation

Recent studies on ZSAS have highlighted the potential of foundation models such as CLIP [38] for localizing anomalies in unseen categories. Most existing methods [10, 11, 19, 31, 35, 36, 56, 58] leverage CLIP's strong vision–language alignment and improve fine-

grained anomaly understanding through feature adaptation or prompt learning. Early works like APRILN-GAN [11] and CLIP-AD [12] introduce learnable linear layers to project image patch features into the joint embedding space. Later methods strengthen cross-category ZSAS performance through various prompt-learning designs, including static learnable prompts (AnomalyCLIP [56]), visual-context prompts (VCP-CLIP [35]), hybrid static–dynamic prompts (AdaCLIP [10]), and prompt distribution learning (Bayes-PFL [36]). Several LMM-based methods have recently been proposed for zero-shot tasks [17, 18, 24, 49], such as Anomaly-OV [49] and IAD-R1 [24]. However, they primarily focus on image-level anomaly understanding and classification rather than segmentation. In contrast to the above methods, this work aims to develop an instruction-following ZSAS model based on LMMs that directly outputs binarized segmentation masks, which is more practical in real industrial and medical scenarios.

3. Method

Following the recent ZSAD setting [36, 56], our model is trained under supervision on seen categories C^s from an auxiliary dataset including our curated Anomaly-Instruct20K, and directly tested on unseen categories C^u collected from industrial and medical domains. Importantly, C^s and C^u are disjoint, and the datasets used for training and evaluation exhibit substantial domain gaps.

3.1. Model Design

Framework overview. As illustrated in Figure 3, AG-VAS is an anchor-guided framework that integrates absolute and relative semantic anchors into an instruction-driven segmentation pipeline. It consists of four key components: 1) a frozen semantic image encoder $\mathcal{F}_s$ and pixel image encoder $\mathcal{F}_p$; 2) a large language model $\mathcal{F}_{\text{LLM}}$; 3) a Semantic–Pixel Alignment Module (SPAM); and 4) a lightweight Anchor-Guided Mask Decoder $\mathcal{D}$.

Given an input image x_{img} and a text query x_{txt}, the frozen encoders first extract semantic features f_s and pixel-level features f_p. These features are processed by the LLM together with the anchor tokens [NOR]/[ANO] and [SEG]. The mask decoder $\mathcal{D}$ then decodes the final-layer embedding corresponding to each anchor token to produce the binary segmentation mask $\mathbf{M}$. SPAM enhances this process by aligning the LLM's semantic representations with the pixel-level features, enabling the anchor embeddings to effectively guide mask generation. In the following, we describe each component in detail.

Design of Semantic Anchors. In LMM-based ZSAS, we define *semantic anchors* as bridges that connects the high-level semantic representations produced by the LMM with the mask decoder. The key motivation is to enable the model to capture fine-grained anomaly-related semantics through specific anchor embeddings in the LMM output space. From the perspective of a human inspector, locating anomalies in unseen categories generally relies on two complementary reasoning processes: 1) leveraging prior *world knowledge* and contextual understanding of common defect types (e.g., holes, cracks, scratches) to recognize what anomalies might look like and where they typically appear; and 2) contrasting candidate regions against surrounding normal areas to identify inconsistencies or irregular patterns.

Inspired by these two reasoning modes, we introduce two types of semantic anchors: the **absolute anchor** [SEG], which encodes world knowledge and contextual cues regarding the appearance, structure, and location of potential anomalies, and the **relative anchors** [NOR]/[ANO], which guide the model to perceive anomalies in relation to normal patterns. These anchors are implemented as special tokens that are inserted into the text sequence and added in the LMM's vocabulary space, allowing them to learn task-specific embeddings that differ from those obtained through conventional prompt tuning. During inference, AG-VAS mainly supports two segmentation modes that differ in whether the reasoning process is explicitly externalized:

(1) *Implicit Mode*. The model directly segments anomalies by leveraging internalized contextual and world knowledge. In this work, this corresponds to the Direct Segmentation scenario. For example:

> ***User:*** *Please segment the anomalies in this image.*
> ***Assistant:*** *Sure, it is [NOR][ANO][SEG].*

(2) *Explicit Mode*. The model externalizes its reasoning process either before or after mask prediction, such as describing anomalies before segmentation (Describe-then-Segment) or explaining the predicted mask (Segment-then-Explain). As an example, in the Describe-then-Segment process:

> ***User:*** *What anomalies are present in this image? Please describe the anomalies and then output the segmentation map.*
> ***Assistant:*** *The anomaly is a small hole located on the upper part of the hazelnut, as indicated by [NOR][ANO][SEG].*

Both modes rely on identical anchor representations and world knowledge grounding; they differ only in whether the description or reasoning process is explicitly articulated. Unless otherwise specified, we adopt the Direct Segmentation mode for quantitative evaluation.

Semantic–Pixel Alignment Module. The introduction of semantic anchors bridges the high-level semantic understanding of LMMs to pixel-level segmentation. However, a gap exists between the high-level semantic embedding space of the LMM and the low-level visual features of the pixel encoder, hindering the anchors' alignment with spatially coherent visual cues. Therefore, we propose the SPAM, which facilitates the alignment of pixel-level image features with the LMM's embedding space through cross-modal interaction.

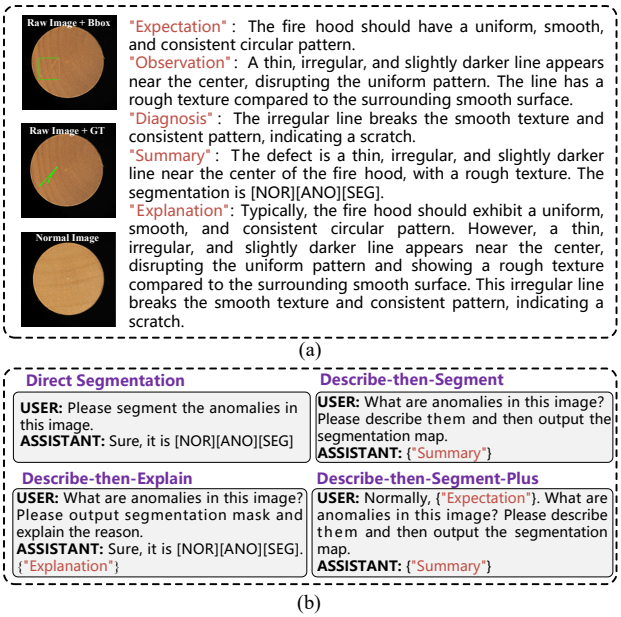


Figure 4. (a) An example from our Anomaly-Instruct20K dataset. (b) An example of instruction-tuning with dynamic fusion of templates during the actual training process. Note that only one specific template is shown here, with further details provided in the Appendix.

Specifically, given an image $x_{\text{img}} \in \mathbb{R}^{3\times h\times w}$, the semantic image encoder and the pixel image encoder are first used separately to extract visual features. The extracted features are then mapped to the same embedding dimension using a simple linear layer:

$$f_s = Linear_s(\mathcal{F}_s(x_{\text{img}})) \quad f_p = Linear_p(\mathcal{F}_p(x_{\text{img}})) \quad (1)$$

where $\mathcal{F}_s$ and $\mathcal{F}_p$ are semantic and pixel encoder, respectively. To facilitate better interaction between the anchors and the mask decoder, we employ a cross-modal attention module, where the semantic features extracted by the semantic image encoder serve as queries $\mathbf{Q}$ to attend to pixel embeddings $(\mathbf{K}, \mathbf{V})$:

$$f_{align} = MHCA(\mathbf{Q} = f_s, \mathbf{K} = f_p, \mathbf{V} = f_p) \quad (2)$$

where $MHCA$ denotes the multi-head cross attention [44]. Subsequently, the image embeddings f_s from the semantic encoder, the aligned embeddings f_{align}, and the text embeddings f_{txt} are concatenated and jointly fed into the large language model, which outputs a text response $\hat{y}_{\text{txt}}$:

$$\hat{y}_{\text{txt}} = \mathcal{F}_{\text{LLM}}(Concat(f_s, f_{align}, f_{\text{txt}})) \quad (3)$$

where f_{txt} is a tokenized version of the input text query x_{txt}. The LLM output $\hat{y}_{\text{txt}}$ contains our designed anchor tokens [NOR],[ANO],[SEG]. We extract the LLM last-layer embeddings corresponding to these tokens and map them into the mask decoder space as $h_{nor}, h_{ano}, h_{seg}$ using a shared Token Refiner $\mathcal{F}_{tr}$, which consists of two linear layers.

Anchor-Guided Mask Decoder. Given the refined anchor embeddings $h_{\text{nor}}, h_{\text{ano}}, h_{\text{seg}}$ and the pixel embeddings f_p, the proposed AGMD aims to generate absolute and relative segmentation masks guided by semantic anchors. We first concatenate the anchor embeddings with three learnable query tokens $\mathbf{T}_{\text{learn}} = \{t_{\text{nor}}, t_{\text{ano}}, t_{\text{seg}}\}$ to form the decoder input:

$$\mathbf{Z}_0 = [t_{\text{nor}}, t_{\text{ano}}, t_{\text{seg}}, h_{\text{nor}}, h_{\text{ano}}, h_{\text{seg}}]. \quad (4)$$

The decoder performs L layers of bidirectional cross-attention between the initial latent feature $\mathbf{Z}_0$ and pixel embeddings f_p, resulting in the final output $\mathbf{Z}_L$ and updated pixel embeddings f'_p. The process is described as:

$$\mathbf{Z}_L, f'_p = \text{BiAttn}^L(\mathbf{Z}_0, f_p) \quad (5)$$

where $\text{BiAttn}^L(\cdot)$ refers to applying a sequence of L bidirectional cross-attention blocks, iteratively updating both $\mathbf{Z}_0$ and f_p, and following the design principles of SAM [20]. After the final layer, we select the updated learnable tokens $\mathbf{T}'_{\text{learn}} = \{t'_{\text{nor}}, t'_{\text{ano}}, t'_{\text{seg}}\}$ from $\mathbf{Z}_L$ to compute segmentation masks. More architectural details of AGMD are provided in the Appendix.

For the absolute semantic anchor [SEG], a foreground probability map is produced via a sigmoid activation $\sigma(\cdot)$:

$$P_{\text{seg}} = \sigma(t'_{\text{seg}} f'^{\top}_p), \quad (6)$$

while for the relative anchors [NOR] and [ANO], a contrastive softmax function is applied to produce normal–abnormal probability maps:

$$[P_{\text{nor}}, P_{\text{ano}}] = \text{Softmax}([t'_{\text{nor}}, t'_{\text{ano}}] f'^{\top}_p). \quad (7)$$

The final anomaly map is obtained by fusing the absolute and relative maps:

$$P = \alpha \cdot P_{\text{seg}} + (1-\alpha) \cdot P_{\text{ano}}, \quad \alpha = 0.5 \quad (8)$$

Finally, the binary segmentation mask $\mathbf{M}$ is produced by thresholding the fused map at 0.5. This design allows the decoder to transform abstract anchor semantics into pixel-level representations under the guidance of both absolute ([SEG]) and relative ([NOR]/[ANO]) anchors. Detailed implementation and architecture specifications of AGMD are provided in the Appendix.

3.2. Training Objectives

The training of AG-VAS involves two complementary objectives:

1) **Textual Autoregressive Loss.** All tokens in the target sequence, including the anchor tokens, are supervised using

the standard autoregressive cross-entropy loss:

$$\mathcal{L}_{\text{txt}} = -\frac{1}{T_y}\sum_{t=1}^{T_y}\log P(y_t \mid y_{<t}, x_{\text{img}}, x_{\text{txt}}) \tag{9}$$

where y_t denotes the ground-truth token at step t, T_y is the length of the target sequence, and x_{img}, x_{text} are the input image and text query, respectively.

2) **Segmentation Loss.** For the predicted masks from the Anchor-Guided Mask Decoder, we unify the loss for all anchors as a combination of Binary Cross-Entropy (BCE) loss and Dice loss:

$$\mathcal{L}_{\text{seg}} = \sum_{c}\Big(\lambda_{bce}\text{BCE}(P_c, M_c) + \lambda_{dic}\text{Dice}(P_c, M_c)\Big) \tag{10}$$

where $c \in \{\text{SEG}, \text{NOR}, \text{ANO}\}$,$P_c$ is the predicted mask for anchor c, and M_c is the corresponding ground-truth mask. For the relative normal anchor, the ground-truth mask is the complement of the abnormal mask. The weights λ_{bce} and λ_{dic} are set to 0.5 and 2.0 empirically.

The total training objective is then defined as:

$$\mathcal{L} = \mathcal{L}_{\text{txt}} + \mathcal{L}_{\text{seg}} \tag{11}$$

3.3. Anomaly-Instruct20K Dataset

To improve the instruction-based segmentation ability of large multimodal models, we introduce a new dataset, Anomaly-Instruct20K, designed to inject world knowledge on normal and abnormal visual patterns into the LMM backbone. Different from prior datasets such as Anomaly-Instruct-125K [49], which mainly provide question–answer annotations for general anomaly understanding, our dataset bridges vision–language reasoning with pixel-level anomaly perception through instruction-driven segmentation supervision. Anomaly-Instruct20K is built using samples from three publicly available datasets: RealIAD [45], GoodsAD [53], and DTD-Synthetic [3].

We generate the dataset with the scientific multimodal model *Inter-S1-241B* [4]. For each anomalous sample, the model receives three visual inputs: 1) the raw image with bounding boxes, 2) the raw image overlaid with the ground-truth mask, and 3) a corresponding normal reference image. Conditioned on these inputs, Inter-S1 generates a structured annotation comprising five fields: *Expectation*, which describes the object's ideal normal appearance; *Observation*, which details concrete visual deviations (e.g., location, shape, size, or texture irregularities); *Diagnosis*, which explains why these deviations break normal material or structural consistency; *Summary*, which converts the detected cues into a concise segmentation instruction and explicitly encodes spatial and appearance attributes necessary for grounding the anchor tokens [NOR][ANO][SEG]; and *Explanation*, which synthesizes all components into a coherent reasoning paragraph. An example of the generated annotations, along with the visual inputs, is shown in Figure 4(a).

During training, we generate final instructions by sampling templates from a library and combining them with dataset fields, covering four main instruction types: Direct Segmentation, Describe-then-Segment, Describe-then-Segment-Plus and Segment-then-Explain. An example of combining a template to generate a real instruction-tuning sample during the training process is shown in Figure 4(b). More detailed dataset construction procedures and system prompts are included in the Appendix.

4. Experiment

4.1. Experimental Setup

Datasets. The training of AG-VAS follows a multi-task joint learning strategy, leveraging datasets from two complementary domains: general object segmentation and anomaly segmentation. For general object segmentation, we adopt the setup from LISA [21], using datasets such as ADE20K [55] to provide comprehensive segmentation capabilities. For anomaly segmentation, we employ two sources of auxiliary training data: 1) the constructed Anomaly-Instruct20K, and 2) a random set of 20k images extracted from existing industrial anomaly segmentation datasets [3, 5, 30, 40, 52], referred to as Anomaly-Seg20K, used under direct segmentation instructions. To preserve the model's dialogue and instruction-following capabilities, we include VQA datasets (LLaVA-150K [28] and WebAD [49]) during training. This joint training enables the absolute semantic anchor [SEG] to align with specific defect entities (e.g., holes, scratches) through contextual semantics and world knowledge, while the relative anchors [NOR] and [ANO] capture the semantic contrast between normal and anomalous regions. To evaluate the ZSAS performance of our model, we follow the protocol in [36] and conduct experiments on six real-world datasets, including three from the industrial domain (MVTec-AD [7], KSDD2 [9], RSDD [51]) and three from the medical domain (ISIC [14], ColonDB [42], ClinicDB [8]). More details about the datasets can be found in Appendix.

Evaluation Metrics. In this work, we focus on the quality of ZSAS predictions, particularly the binarized segmentation accuracy. We report pixel-level average precision (AP), the maximum F1 score at the optimal threshold (F1-Max), and the Intersection over Union of anomalous samples (IoU_{ano}). To further evaluate the model's ability to reject normal images, we define IoU_{nor}: the model receives a score of 1 if it predicts an empty mask for a normal sample, and 0 otherwise. Both IoU_{ano} and IoU_{nor} are computed at the per-image level and then averaged over the correspond-

Table 1. Comparison with existing state-of-the-art methods. The best results are marked in red, while the second-best methods are indicated in blue. The superscript * indicates that these methods were retrained on the anomaly segmentation dataset used in our AG-VAS.

	Method	Base Model	Industrial Datasets			Medical Datasets		
			MVTec-AD	KSDD2	RSDD	ISIC	ColonDB	ClinicDB
CLIP-Based (AP, F1-Max, IoU_{ano})	WinCLIP [19]	CLIP(ViT-B-16-240)	(18.0, 24.3, 0.0)	(17.1, 24.6, 0.0)	(2.1, 6.6, 0.1)	(62.4, 64.1, 0.24)	(14.3, 21.0, 0.0)	(19.4, 27.3, 0.0)
	APRIL-GAN [11]	CLIP(ViT-L-14-336)	(40.8, 43.3, 25.1)	(61.6, 62.6, 41.8)	(30.6, 40.2, 24.3)	(69.8, 69.3, 12.8)	(23.2, 32.5, 14.4)	(38.8, 44.4, 25.3)
	AnomalyCLIP [56]	CLIP(ViT-L-14-336)	(34.5, 39.1, 22.6)	(41.8, 50.3, 27.6)	(19.1, 30.6, 8.8)	(74.4, 70.8, 51.0)	(31.3, 39.8, 25.2)	(42.2, 45.9, 30.0)
	AnomalyCLIP* [56]	CLIP(ViT-L-14-336)	(35.2, 39.0, 22.4)	(48.8, 53.6, 28.6)	(25.5, 34.4, 8.5)	(76.5, 71.4, 54.1)	(34.2, 39.8, 22.7)	(44.1, 47.3, 29.4)
	Bayes-PFL [36]	CLIP(ViT-L-14-336)	(48.3, 49.1, 22.2)	(73.7, 63.5, 25.4)	(39.1, 45.2, 3.7)	(84.6, 76.8, 60.0)	(31.9, 39.2, 16.8)	(53.2, 51.7, 21.0)
	Bayes-PFL* [36]	CLIP(ViT-L-14-336)	(50.3, 50.4, 29.9)	(74.9, 67.6, 39.5)	(41.9, 46.9, 7.7)	(81.9, 74.3, 52.1)	(30.5, 38.1, 27.7)	(47.6, 50.7, 34.4)
LMM-Based (AP, F1-Max, IoU_{ano})	PixelLM [39]	LLaVA-13B	(13.6, 19.6, 11.9)	(42.6, 46.8, 32.4)	(5.3, 9.4, 4.9)	(69.8, 61.7, 48.9)	(8.9, 16.8, 9.8)	(13.0, 20.9, 14.6)
	PaDT [41]	Qwen2.5-VL-7B	(6.8, 12.9, 16.4)	(1.2, 2.9, 12.4)	(0.9, 1.8, 1.7)	(65.9, 68.3, 54.5)	(35.4, 48.4, 39.1)	(39.8, 51.4, 49.8)
	LISA [21]	LLaVA-7B	(9.8, 16.9, 10.2)	(19.5, 23.8, 18.2)	(6.9, 12.5, 2.9)	(68.1, 66.3, 54.6)	(25.9, 33.1, 19.4)	(50.0, 45.3, 36.4)
	LISA [21]	LLaVA-13B	(14.6, 21.8, 13.7)	(36.8, 39.8, 29.4)	(5.2, 11.4, 2.6)	(87.1, 78.1, 64.8)	(35.3, 34.0, 26.2)	(56.3, 52.9, 42.3)
	LISA* [21]	LLaVA-13B	(37.6, 40.8, 31.8)	(63.7, 63.8, 50.9)	(52.1, 57.1, 35.2)	(88.7, 82.4, 64.1)	(40.9, 43.2, 29.5)	(67.8, 62.7, 48.3)
	LISA [21]	LLaVA-OneVision-7B	(32.4, 38.3, 32.2)	(59.8, 65.8, 49.4)	(31.6, 37.6, 22.6)	(90.7, 81.6, 56.9)	(29.1, 35.8, 14.4)	(56.1, 54.4, 32.6)
	LISA* [21]	LLaVA-OneVision-7B	(41.0, 44.1, 32.3)	(69.4, 69.8, 48.4)	(48.8, 54.9, 34.6)	(84.9, 77.6, 50.1)	(67.2, 62.9, 43.8)	(81.0, 74.1, 59.8)
	AG-VAS	LLaVA-13B	(40.5, 42.5, 37.0)	(66.5, 65.4, 53.7)	(54.3, 57.0, 35.9)	(88.8, 82.7, 64.6)	(43.7, 45.1, 32.2)	(68.3, 65.5, 50.1)
	AG-VAS	LLaVA-OneVision-7B	(51.0, 52.7, 44.8)	(76.6, 74.7, 53.8)	(56.2, 57.8, 36.9)	(92.7, 84.9, 65.0)	(70.7, 66.2, 58.2)	(86.6, 79.2, 69.5)

Table 2. Comparison of the model's ability to reject segmentation on normal samples on the MVTec-AD.

Method	Base Model	IoU_{ano}	IoU_{nor}	Average
PixelLM [39]	LLaVA-13B	11.9	2.4	7.2
PaDT [41]	Qwen2.5-VL-7B	16.4	0.0	8.2
LISA [21]	LLaVA-7B	10.2	1.9	6.1
LISA [21]	LLaVA-13B	13.7	6.3	10.0
LISA* [21]	LLaVA-13B	31.8	43.5	37.7
LISA [21]	LLaVA-OneVision-7B	32.2	4.0	18.1
LISA* [21]	LLaVA-OneVision-7B	32.3	80.9	56.6
AG-VAS	LLaVA-OneVision-7B	**44.8**	**87.7**	**66.3**

ing subsets.

Implementation Details. By default, we adopt LLaVA-OneVision-7B [22] as our LMM backbone, equipped with a semantic image encoder (siglip-patch14-384), while SAM-ViT-H serves as the pixel image encoder for mask prediction. Training is implemented using DeepSpeed with the AdamW optimizer at a learning rate of 0.0003. During instruction tuning, the LLM is fine-tuned with LoRA at rank 16, using a global batch size of 80. The full training of the 7B model takes about 30 hours on four A100 GPUs with 40 GB memory each. More implementation details can be found in the Appendix.

4.2. Comparison with State-of-the-art methods

Since LMM-based ZSAS is still in an early exploratory stage, directly comparable approaches are limited. Therefore, we include four representative CLIP-based ZSAS methods (WinCLIP [19], APRIL-GAN [11], AnomalyCLIP [56], and Bayes-PFL [36]) together with three segmentation methods that address referring segmentation tasks, specifically designed for natural scene understanding: PixelLM [39], PaDT [41], and LISA [21]. To ensure a fair comparison, we also retrained other methods (e.g., AnomalyCLIP, Bayes-PFL, and LISA) on the anomaly segmentation dataset used by our AG-VAS during training. More details regarding the comparison of these methods are shown in Appendix.

Quantitative Comparison. Table 1 presents a comparison of our AG-VAS with other state-of-the-art methods across various industrial and medical datasets. It can be observed that when using the default base model, LLaVA-OneVision, our ZSAS performance surpasses all other methods on every metric. CLIP-based methods, such as Bayes-PFL, achieve comparable AP and F1-Max scores relative to LMM-based approaches but perform poorly on metric IoU_{ano}, which measures true segmentation accuracy. This suggests that their anomaly localization results are ambiguous, leading to lower IoU_{ano} after binarization. Furthermore, we observe that LLM-based methods demonstrate superior zero-shot generalization on medical datasets compared to CLIP-based approaches, particularly after fine-tuning on our carefully curated Anomaly-Instruct20K and Anomaly-Seg20K datasets. It is noteworthy that no medical images were included in the training dataset, demonstrating that LMMs can leverage pre-existing world knowledge for zero-shot generalization to unseen domains. Furthermore, the introduction of the anchor mechanism allows our method to more effectively align semantic and pixel-level features, leading to substantial improvements over other LMMs in leveraging contextual semantics during training.

Comparison of rejection performance. For LMM-based ZSAS methods, it is crucial to not only accurately segment anomalous regions but also to reject normal samples by predicting empty masks, even when the user provides instructions to segment anomalies. The experimental results in Table 2 show that most LMMs struggle in this regard, whereas our AG-VAS demonstrates strong rejection ability, achieving an IoU_{nor} of 87.7% on normal samples.

Table 3. Ablation on different components.

		AP	F1-Max	IoU_{nor}	IoU_{ano}
Module Ablation	w/o [SEG]	49.1	50.9	85.6	42.1
	w/o [NOR][ANO]	47.2	49.7	52.1	39.5
	w/o SPAM	46.5	48.0	70.6	41.4
Training Dataset Ablation	w/o Anomaly-Instruct20K	48.4	50.0	85.2	39.9
	w/o Anomaly-Seg20K	49.3	51.5	53.5	42.4
	w/o General Segmentation	36.1	36.5	70.2	34.9
	AG-VAS	**51.0**	**52.7**	**87.7**	**44.8**

Table 4. Ablation on different segmentation mode during inference.

Segmentation Mode	AP	F1-Max	IoU_{nor}	IoU_{ano}
Describe-then-segment	49.3	51.3	58.8	43.6
Describe-then-segment-plus	**51.9**	**53.0**	90.2	**45.1**
Segment-then-explain	48.6	51.4	**91.1**	43.4
Direct Segmentation	51.0	52.7	87.7	44.8

This indicates a substantial reduction in over-segmentation risk, while AG-VAS also maintains superior performance on anomalous regions, achieving an IoU_{ano} of 44.8%.

4.3. Ablation

In this subsection, ablation experiments on the MVTec-AD dataset are conducted to further investigate the impact of different settings on the proposed AG-VAS.

Influence of Different Components. Table 3 presents an ablation study evaluating the impact of various modules and training datasets on ZSAS performance. From the module ablations, it can be observed that removing the absolute semantic anchor [SEG] leads to a noticeable drop in all metrics, highlighting its crucial role in aligning anchor semantics with concrete defect entities. Excluding the relative anchors [NOR] and [ANO] significantly decreases IoU_{nor}, indicating their importance in capturing the semantic contrast between normal and anomalous regions. The removal of SPAM also causes performance degradation, suggesting its effectiveness in feature aggregation.

Regarding dataset ablations, training without Anomaly-Instruct20K leads to a noticeable decrease in performance, particularly on metrics reflecting anomaly segmentation. This indicates that instruction tuning with Anomaly-Instruct20K successfully injects domain-specific world knowledge into the model, enabling the anchors to better interpret and localize anomalies. Excluding Anomaly-Seg20K causes a larger drop in IoU_{nor}, showing that these auxiliary images are crucial for learning robust normal–abnormal contrasts. Omitting general segmentation data results in the most substantial decrease across all metrics, confirming that multi-domain joint training is essential for aligning absolute semantic anchors with visual features.

Influence of different segmentation modes during inference. Table 4 compares AG-VAS performance under different inference strategies. *Describe-then-Segment* refers to first generating a textual description of the target anomaly and then producing a segmentation mask. *Describe-then-Segment-Plus* prepends an additional sentence describing the normal appearance of the object to the input instruction before the anomaly description. This serves as an expert-provided prior, explicitly informing the model about the object's typical state. *Segment-then-Explain* first predicts the mask and then generates an explanation, while *Direct Segmentation* predicts the mask directly from the instruction without intermediate text.

Compared with the default *Direct Segmentation* mode, *Describe-then-Segment* and *Segment-then-Explain* slightly reduce segmentation performance due to textual noise introduced during intermediate reasoning. However, these modes enable the model to better follow user instructions and support interactive reasoning, justifying a minor sacrifice in mask accuracy. Notably, *Describe-then-Segment-Plus*, by providing explicit knowledge of the object's normal state, improves contextual understanding and partially mitigates the performance drop, leading to the best overall results among the descriptive modes.

5. Conclusion

This paper introduces AG-VAS, an anchor-guided framework for zero-shot visual anomaly segmentation that endows large multimodal models with learnable semantic anchors—absolute token [SEG] and relative tokens [NOR]/[ANO]—to jointly encode world knowledge about defect appearance, shape and location while simultaneously modeling the contextual opposition between normal and abnormal regions. By fusing high-level vision–language embeddings with fine-grained pixel features through the Semantic–Pixel Alignment Module, and decoding anchor-consistent masks via the bidirectional Anchor-Guided Mask Decoder, AG-VAS overcomes the lack of structured anomaly semantics and directly outputs accurate binary masks without requiring category-specific retraining. Extensive experiments on six challenging industrial and medical benchmarks demonstrate that AG-VAS establishes a new state-of-the-art, significantly outperforming both CLIP-based and recent LMM-based competitors. It shows strong rejection performance on normal images and demonstrates robust generalization across unseen domains, providing a practical solution for real-world zero-shot anomaly segmentation task.

Acknowledgement

This work was supported by the Shandong Provincial Natural Science Foundation (No. ZR2025LGY002), the National Natural Science Foundation of China (No. 62373350), the Youth Innovation Promotion Association CAS (No. 2023145) and the Beijing Nova Program (No. 20240484687).

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

[2] Syed Muhammad Anwar, Muhammad Majid, Adnan Qayyum, Muhammad Awais, Majdi Alnowami, and Muhammad Khurram Khan. Medical image analysis using convolutional neural networks: a review. *Journal of medical systems*, 42:1–13, 2018.

[3] Toshimichi Aota, Lloyd Teh Tzer Tong, and Takayuki Okatani. Zero-shot versus many-shot: Unsupervised texture anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5564–5572, 2023.

[4] Lei Bai, Zhongrui Cai, Yuhang Cao, Maosong Cao, Weihan Cao, Chiyu Chen, Haojiong Chen, Kai Chen, Pengcheng Chen, Ying Chen, et al. Intern-s1: A scientific multimodal foundation model. *arXiv preprint arXiv:2508.15763*, 2025.

[5] Tianpeng Bao, Jiadong Chen, Wei Li, Xiang Wang, Jingjing Fei, Liwei Wu, Rui Zhao, and Ye Zheng. Miad: A maintenance inspection dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 993–1002, 2023.

[6] Xiaoyi Bao, Siyang Sun, Shuailei Ma, Kecheng Zheng, Yuxin Guo, Guosheng Zhao, Yun Zheng, and Xingang Wang. Cores: Orchestrating the dance of reasoning and segmentation. In *European Conference on Computer Vision*, pages 187–204. Springer, 2024.

[7] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9592–9600, 2019.

[8] Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized medical imaging and graphics*, 43:99–111, 2015.

[9] Jakob Božič, Domen Tabernik, and Danijel Skočaj. Mixed supervision for surface-defect detection: From weakly to fully supervised learning. *Computers in Industry*, 129: 103459, 2021.

[10] Yunkang Cao, Jiangning Zhang, Luca Frittoli, Yuqi Cheng, Weiming Shen, and Giacomo Boracchi. Adaclip: Adapting clip with hybrid learnable prompts for zero-shot anomaly detection. In *European Conference on Computer Vision*, pages 55–72. Springer, 2025.

[11] Xuhai Chen, Yue Han, and Jiangning Zhang. A zero-/few-shot anomaly classification and segmentation method for cvpr 2023 vand workshop challenge tracks 1&2: 1st place on zero-shot ad and 4th place on few-shot ad. *arXiv preprint arXiv:2305.17382*, 2023.

[12] Xuhai Chen, Jiangning Zhang, Guanzhong Tian, Haoyang He, Wuhao Zhang, Yabiao Wang, Chengjie Wang, and Yong Liu. Clip-ad: A language-guided staged dual-path model for zero-shot anomaly detection. In *International Joint Conference on Artificial Intelligence*, pages 17–33. Springer, 2024.

[13] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024.

[14] Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 168–172. IEEE, 2018.

[15] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023.

[16] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023.

[17] Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. Anomalygpt: Detecting industrial anomalies using large vision-language models. In *Proceedings of the AAAI conference on artificial intelligence*, pages 1932–1940, 2024.

[18] Wei Guan, Jun Lan, Jian Cao, Hao Tan, Huijia Zhu, and Weiqiang Wang. Emit: Enhancing mllms for industrial anomaly detection via difficulty-aware grpo. *arXiv preprint arXiv:2507.21619*, 2025.

[19] Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. Winclip: Zero-/few-shot anomaly classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19606–19616, 2023.

[20] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

[21] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9579–9589, 2024.

[22] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[23] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[24] Yanhui Li, Yunkang Cao, Chengliang Liu, Yuan Xiong, Xinghui Dong, and Chao Huang. Iad-r1: Reinforcing consistent reasoning in industrial anomaly detection. *arXiv preprint arXiv:2508.09178*, 2025.

[25] Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. Monkey: Image resolution and text label are important things for large multi-modal models. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26763–26773, 2024.

[26] Zhang Li, Biao Yang, Qiang Liu, Shuo Zhang, Zhiyin Ma, Liang Yin, Linger Deng, Yabo Sun, Yuliang Liu, and Xiang Bai. Lira: Inferring segmentation in large multi-modal models with local interleaved region assistance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 24056–24067, 2025.

[27] Bin Lin, Zhenyu Tang, Yang Ye, Jiaxi Cui, Bin Zhu, Peng Jin, Jinfa Huang, Junwu Zhang, Yatian Pang, Munan Ning, et al. Moe-llava: Mixture of experts for large vision-language models. *arXiv preprint arXiv:2401.15947*, 2024.

[28] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[29] Zuyan Liu, Yuhao Dong, Yongming Rao, Jie Zhou, and Jiwen Lu. Chain-of-spot: Interactive reasoning improves large vision-language models. *arXiv preprint arXiv:2403.12966*, 2024.

[30] Xiaoming Lv, Fajie Duan, Jia-jia Jiang, Xiao Fu, and Lin Gan. Deep metallic surface defect detection: The new benchmark and detection network. *Sensors*, 20(6):1562, 2020.

[31] Wenxin Ma, Xu Zhang, Qingsong Yao, Fenghe Tang, Chenxu Wu, Yingtai Li, Rui Yan, Zihang Jiang, and S Kevin Zhou. Aa-clip: Enhancing zero-shot anomaly detection via anomaly-aware clip. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4744–4754, 2025.

[32] Ji Qi, Ming Ding, Weihan Wang, Yushi Bai, Qingsong Lv, Wenyi Hong, Bin Xu, Lei Hou, Juanzi Li, Yuxiao Dong, et al. Cogcom: Train large vision-language models diving into details through chain of manipulations. 2024.

[33] Shichen Qu, Xian Tao, Xinyi Gong, Zhen Qu, Mukesh Prasad, Fei Shen, Zhengtao Zhang, and Guiguang Ding. Lscad: A large-small model collaboration framework for unsupervised industrial anomaly detection. *IEEE Transactions on Instrumentation and Measurement*, 2025.

[34] Zhen Qu, Xian Tao, Fei Shen, Zhengtao Zhang, and Tao Li. Investigating shift equivalence of convolutional neural networks in industrial defect segmentation. *IEEE Transactions on Instrumentation and Measurement*, 72:1–17, 2023.

[35] Zhen Qu, Xian Tao, Mukesh Prasad, Fei Shen, Zhengtao Zhang, Xinyi Gong, and Guiguang Ding. Vcp-clip: A visual context prompting model for zero-shot anomaly segmentation. In *European Conference on Computer Vision*, pages 301–317. Springer, 2024.

[36] Zhen Qu, Xian Tao, Xinyi Gong, Shichen Qu, Qiyu Chen, Zhengtao Zhang, Xingang Wang, and Guiguang Ding. Bayesian prompt flow learning for zero-shot anomaly detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 30398–30408, 2025.

[37] Zhen Qu, Xian Tao, Xinyi Gong, ShiChen Qu, Xiaopei Zhang, Xingang Wang, Fei Shen, Zhengtao Zhang, Mukesh Prasad, and Guiguang Ding. Dictas: A framework for class-generalizable few-shot anomaly segmentation via dictionary lookup. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20519–20528, 2025.

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

[39] Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. Pixellm: Pixel reasoning with large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26374–26383, 2024.

[40] Yong Shi, Limeng Cui, Zhiquan Qi, Fan Meng, and Zhensong Chen. Automatic road crack detection using random structured forests. *IEEE Transactions on Intelligent Transportation Systems*, 17(12):3434–3445, 2016.

[41] Yongyi Su, Haojie Zhang, Shijie Li, Nanqing Liu, Jingyi Liao, Junyi Pan, Yuan Liu, Xiaofen Xing, Chong Sun, Chen Li, et al. Patch-as-decodable-token: Towards unified multi-modal vision tasks in mllms. *arXiv preprint arXiv:2510.01954*, 2025.

[42] Nima Tajbakhsh, Suryakanth R Gurudu, and Jianming Liang. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE transactions on medical imaging*, 35(2):630–644, 2015.

[43] Xian Tao, Xinyi Gong, Xin Zhang, Shaohua Yan, and Chandranath Adak. Deep learning for unsupervised anomaly localization in industrial images: A survey. *IEEE Transactions on Instrumentation and Measurement*, 71:1–21, 2022.

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[45] Chengjie Wang, Wenbing Zhu, Bin-Bin Gao, Zhenye Gan, Jiangning Zhang, Zhihao Gu, Shuguang Qian, Mingang Chen, and Lizhuang Ma. Real-iad: A real-world multi-view

dataset for benchmarking versatile industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22883–22892, 2024.

[46] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[47] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Song XiXuan, et al. Cogvlm: Visual expert for pretrained language models. *Advances in Neural Information Processing Systems*, 37:121475–121499, 2024.

[48] Zhuofan Xia, Dongchen Han, Yizeng Han, Xuran Pan, Shiji Song, and Gao Huang. Gsva: Generalized segmentation via multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3858–3869, 2024.

[49] Jiacong Xu, Shao-Yuan Lo, Bardia Safaei, Vishal M Patel, and Isht Dwivedi. Towards zero-shot anomaly detection and reasoning with multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 20370–20382, 2025.

[50] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 13040–13051, 2024.

[51] Haomin Yu, Qingyong Li, Yunqiang Tan, Jinrui Gan, Jianzhu Wang, Yangli-ao Geng, and Lei Jia. A coarse-to-fine model for rail surface defect detection. *IEEE Transactions on Instrumentation and Measurement*, 68(3):656–666, 2018.

[52] Chenkai Zhang, Shaozhe Feng, Xulongqi Wang, and Yueming Wang. Zju-leaper: A benchmark dataset for fabric defect detection and a comparative study. *IEEE Transactions on Artificial Intelligence*, 1(3):219–232, 2021.

[53] Jian Zhang, Runwei Ding, Miaoju Ban, and Linhui Dai. Pku-goodsad: A supermarket goods dataset for unsupervised anomaly detection and segmentation. *IEEE Robotics and Automation Letters*, 9(3):2008–2015, 2024.

[54] Tao Zhang, Xiangtai Li, Hao Fei, Haobo Yuan, Shengqiong Wu, Shunping Ji, Chen Change Loy, and Shuicheng Yan. Omg-llava: Bridging image-level, object-level, pixel-level reasoning and understanding. *Advances in neural information processing systems*, 37:71737–71767, 2024.

[55] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017.

[56] Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. In *The Twelfth International Conference on Learning Representations*, 2023.

[57] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

[58] Jiawen Zhu, Yew-Soon Ong, Chunhua Shen, and Guansong Pang. Fine-grained abnormality prompt learning for zero-shot anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22241–22251, 2025.