# From Foundation to Field: LISA Fine-Tuning for Mine Open-Vocabulary Segmentation

JiBo Wang[1], Libin Jiao[1], Zhen Bao[2,3,4], Wenchao Gao[1,*] and Lianzhi Huo[5,*]
{sqt2410405036@student.cumtb.edu.cn, jiaolibin@cumtb.edu.cn, 12079985@ceic.com, gaowc@cumtb.edu.cn, huolz@aircas.ac.cn}

[1]School of Artificial Intelligence, China University of Mining and Technology-Beijing, Beijing, China
[2]CHN Energy Science and Technology and Environment Co., Ltd., China, CHN Energy Zhi Shen Control Technology Co., Ltd., China
[3]State R&D Center of Control System and Information Security Technologies for Energy Industry, China
[4]Beijing Engineering Research Center of Power Station Automation, Beijing, 102211, China
[5]the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China

**Abstract.** Underground mining environments are characterized by dim illumination, cluttered man-made structures, and frequent occlusions, yet most existing underground segmentation methods still treat the task as closed-set pixel classification over a fixed label set, decoupled from natural-language descriptions and unable to dynamically segment context-specific targets according to operator instructions. In this work, we present MineLISA, an instruction-guided segmentation framework adapted from the Language Instructed Segmentation Assistant (LISA) and optimized for industrial underground mining applications. MineLISA takes natural-language prompts as input and produces pixel-level masks for underground mining objects on the MUSeg multimodal semantic-segmentation dataset of underground imagery. To adapt LISA to this domain under realistic resource constraints, we employ a LoRA-based parameter-efficient fine-tuning strategy on the existing vision–language alignment modules and lightweight segmentation decoder, and re-weight the segmentation loss to emphasize thin, safety-critical structures such as cables and pipelines. This design enables MineLISA to better align natural-language instructions with underground visual patterns while remaining suitable for hardware with limited GPU memory. Experiments on MUSeg show that, compared with the original LISA, MineLISA achieves significantly improves instruction-conditioned mask predictions and more stable, generalizable segmentation across diverse underground object categories, indicating strong potential for real-world deployment in coal-mine operations.

**Keywords:** Multimodal large models, open-vocabulary semantic segmentation, underground coal-mine applications, LoRA fine-tuning

IIKI 2025, December 18-21, Chengdu, People's Republic of China

DOI 10.4108/eai.18-12-2025.2365260

## 1 Introduction

Underground coal-mine environments are highly complex: low illumination and flickering lights reduce contrast and amplify noise; dust and moisture introduce strong scattering and occlusion; pipelines, cables, and ventilation ducts often exhibit thin structures and large geometric variability. Meanwhile, operational demands are highly dynamic: instead of merely segmenting a fixed set of semantic categories, operators must reason about the current production context and focus on situation-specific targets such as a leaking pipe, an exposed cable, or a newly installed sign[1, 2]. Image-based visual perception—especially semantic segmentation—has therefore become central to hazard screening and target identification in underground mines. However, conventional semantic segmentation models treat segmentation as pixel-wise classification over a fixed label ontology and do not accept natural-language inputs[3]. As a result, they are fundamentally closed-set: such models can only predict a predefined set of categories and cannot leverage task-specific descriptions or contextual cues to perform reasoning-aware, instruction-following, on-demand segmentation of arbitrary underground targets. These limitations are further exacerbated by the limited compute and GPU memory (VRAM) typically available in industrial deployments.

These observations lead to the following research question: how can a segmentation system under realistic resource constraints be designed to perform reasoning-aware, instruction-following segmentation on underground coal-mine imagery? To address this question, we investigate whether multimodal foundation models can be leveraged as a resource-efficient backbone for reasoning-driven, instruction-guided segmentation in underground mining environments, moving beyond conventional closed-set pixel classification pipelines.

In this paper, we propose MineLISA, a framework tailored for instruction-guided segmentation in challenging underground mining scenes. MineLISA adopts LISA as its base model and realizes this idea through LISA's architecture: it builds upon a LLaVA-style vision–language backbone and a SAM-based segmentation head, which are connected via an embedding-as-mask interface. Specifically, LISA introduces a special `<SEG>` token whose hidden state, produced by the multimodal backbone, is fed into the SAM-style decoder and transformed into an instruction-aligned mask, thereby effectively bridging natural-language instructions and pixel-level predictions [4]. Building on this paradigm, MineLISA inherits LISA's multimodal backbone and `<SEG>`-based interface, and further specializes them for underground mining scenes. Concretely, we adapt LISA to the MUSeg dataset [5] of industrial underground imagery using a parameter-efficient fine-tuning strategy: the visual backbone is kept frozen, Low-Rank Adaptation (LoRA) [6] is inserted into the vision–language alignment modules to better couple textual instructions with underground visual patterns, and the segmentation loss is re-weighted to emphasize thin, safety-critical structures such as cables and pipelines. This design enables MineLISA to capture domain-specific semantics—such as equipment, pipelines, and signage—from limited image–text pairs while retaining the base model's open-vocabulary and instruction-following capabilities. Experiments on MUSeg show that, compared with the original LISA baseline, MineLISA consistently improves instruction-guided mask quality under low illumination, heavy occlusion, and cluttered layouts, demonstrating its potential as a practical visual reasoning module for autonomous coal-mine operations. Our code, models, and data are available at `https://github.com/jeb223/MineLISA`.

## 2 Related Work

Image segmentation has progressed from classical pixel classification to instance and open/generalized segmentation.Early fully convolutional architectures such as FCN and U-Net established encoder–decoder style semantic segmentation [3, 7]. Subsequent works (e.g., DeepLab) further introduced multi-scale context aggregation and dilated convolutions [8, 9, 10], while two-stage approaches like Mask R-CNN strengthened instance-level mask prediction [11].More recently, SAM introduced promptable segmentation to general settings, performing zero-shot segmentation in new scenes using diverse prompts (points, boxes, existing masks) and showing strong generalization. With the rise of large language models, systems such as LISA combine multimodal LLMs with an "embedding-as-mask" paradigm, using a special token to trigger mask generation and enabling natural-language instruction for complex semantics. LISA produces masks from implicit, compositional text queries, addressing limitations of perception systems that rely on explicit commands and lack reasoning ability[4]. This direction offers both transferability and interactivity, making it promising for real-world deployment.

In mining vision, closed-set labeling with conventional segmentation networks (e.g., U-Net) remains common for roadway facility monitoring, conveyor defect detection, and rock boundary delineation [12]. Some studies introduce more sophisticated preprocessing pipelines or adapt general-purpose models such as SAM and Mask R-CNN to industrial scenarios to improve robustness and zero-shot performance [13]. However, these methods still struggle with open vocabularies and complex semantics involving multiple constraints and compositional object parts. Motivated by these challenges, we adopt a parameter-efficient fine-tuning strategy for LISA on the MuSeg industrial coal-mine dataset, aiming to enable robust instruction-guided mask prediction from complex textual prompts while keeping deployment costs manageable in underground environments.

## 3 Method

Our goal is to transition from general-purpose "understanding instructions" to "robustly segmenting as instructed" in extreme underground environments. To achieve this under strict hardware constraints, we propose a LoRA-tuned architecture based on LISA. As illustrated in Fig. 1, our framework integrates a multimodal large language model (LLM) for reasoning and a specialized vision decoder for precise segmentation. MineLISA extends a large language model (LLM) to image segmentation: it takes an image and a textual prompt as input and outputs the corresponding mask. It follows an "instruction $\rightarrow$ mask" paradigm—encoding the image into visual tokens aligned with the text space. When the special token `[SEG]` appears in the input, the model uses its hidden state as a query and, through a lightweight segmentation head, produces a pixel-level mask ("embedding-as-mask").

Our model follows a "reasoning-as-segmentation" paradigm and consists of three key components. First, a dual-vision encoder is used to balance high-level semantic understanding and low-level spatial precision: a frozen CLIP ViT-L/14 backbone extracts semantic features aligned with the text space, while a frozen SAM ViT-H backbone provides high-resolution spatial embeddings needed to delineate fine structures such as cables and pipes. Second, a multimodal LLM backbone

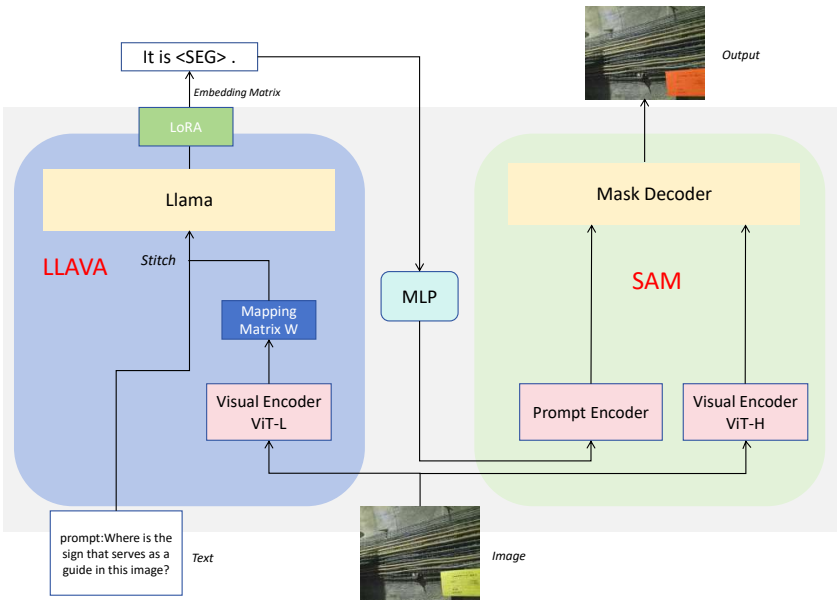


**Fig. 1.** The overall architecture of our proposed method. The system integrates LLaVA for multimodal reasoning and SAM for segmentation. The image is processed by a frozen ViT-L (for semantic understanding) and a ViT-H (for spatial details). The LLM (Llama) processes the user prompt and visual tokens to generate a <SEG> token. This token is projected via an MLP to prompt the SAM decoder, enabling precise segmentation of the target object specified in the text.

initialized from LISA-7B serves as the core reasoning module, taking as input the semantic visual tokens together with the user's textual instruction (e.g., "Where is the protective device in the image?") and producing a natural-language response; crucially, when segmentation is required, the LLM appends a special token [SEG] at the end of its output sequence. Third, in the mask decoding stage, the hidden state of the [SEG] token, $h_{\text{seg}}$, is treated as a dynamic semantic query, projected through an MLP and fed into the SAM mask decoder, which interacts with the spatial features from ViT-H to generate the final pixel-wise binary mask, effectively translating the LLM's intent into a concrete segmentation map.

Directly training a 7B-parameter LLM together with its vision backbones is impractical for edge devices deployed in underground mines. Therefore, we freeze the entire vision encoder (ViT) and most of the LLM weights $W_0$. We then perform fine-tuning by applying low-rank adaptation (LoRA) only to the attention layers of the LLM, specifically to the query ($q_{\text{proj}}$) and value ($v_{\text{proj}}$) projection matrices. This design preserves the pretrained backbone's generalization while focusing capacity on vision–language alignment and mask decoding for coal-mine scenarios. Specifically, given a frozen pretrained weight $W_0 \in \mathbb{R}^{d\times k}$, LoRA introduces a trainable low-rank update $\Delta W$:

$$W = W_0 + \Delta W, \qquad \Delta W = \frac{\alpha}{r} BA, \tag{1}$$

where $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$ are learnable factors, $r$ is the rank (we set $r$=8), and $\alpha$ is a scaling factor (we set $\alpha$=16). We apply (1) only to `q_proj`/`v_proj`, leaving other base LLM weights frozen; a dropout of 0.05 is used on LoRA adapters. The matrix $A$ is initialized with a random Gaussian distribution, while $B$ is initialized to zero, ensuring that $\Delta W = 0$ at the beginning of training. This strategy allows the model to learn mine-specific semantic information (e.g., distinguishing between "cable" and "pipe" under low-light conditions) while modifying only a small number of parameters.

To train the model effectively on the MUSeg dataset—which exhibits severe foreground–background imbalance and thin, elongated structures (such as pipes and support beams)—we design a hybrid loss function. The total loss $\mathscr{L}_{\text{total}}$ is a weighted sum of three components for mask prediction: a cross-entropy loss ($\mathscr{L}_{\text{ce}}$), a binary cross-entropy loss ($\mathscr{L}_{\text{bce}}$), and a Dice loss ($\mathscr{L}_{\text{dice}}$).The overall objective function is formulated as:

$$\mathscr{L}_{\text{total}} \;=\; \lambda_{\text{ce}}\,\mathscr{L}_{\text{ce}} \;+\; \lambda_{\text{dice}}\,\mathscr{L}_{\text{dice}} \;+\; \lambda_{\text{bce}}\,\mathscr{L}_{\text{bce}}, \tag{2}$$

with $(\lambda_{\text{ce}}, \lambda_{\text{dice}}, \lambda_{\text{bce}}) = (0.5, 1.5, 1.0)$, which stabilizes optimization under class imbalance and thin/fragmented structures.For $C$ classes (foreground/background here), with one-hot label $y_c$ and predicted probability $\hat{y}_c$,

$$\mathscr{L}_{\text{ce}} \;=\; -\sum_{c=1}^{C} y_c \log \hat{y}_c. \tag{3}$$

Given predicted probabilities $\mathbf{p} \in [0,1]^N$ and binary ground-truth mask $\mathbf{g} \in \{0,1\}^N$ over $N$ pixels,

$$\mathscr{L}_{\text{dice}} \;=\; 1 \;-\; \frac{2\sum_{i=1}^{N} p_i g_i + \varepsilon}{\sum_{i=1}^{N} p_i^2 \;+\; \sum_{i=1}^{N} g_i^2 \;+\; \varepsilon}, \tag{4}$$

where $\varepsilon$ is a small constant for numerical stability.The pixel-wise BCE term is

$$\mathscr{L}_{\text{bce}} \;=\; -\frac{1}{N}\sum_{i=1}^{N} \Big( g_i \log p_i \;+\; (1-g_i)\log(1-p_i) \Big). \tag{5}$$

In terms of training data processing, we use an instruction-style short template for supervision (e.g., `USER: <image>\n Please segment the {class_name} in this image. ASSISTANT: [SEG].`), where `[SEG]` is a teaching trigger token that aligns textual outputs with mask supervision. If an image is paired with multiple prompts, we expand them into separate samples to ensure a one-to-one mapping of "one instruction $\leftrightarrow$ one image $\leftrightarrow$ one mask." The dialogue template retains exactly one image placeholder, optionally wrapped by unified start/end markers when needed, to avoid feature mismatch from multiple placeholders. On the text side, sequences are encoded under a multi-turn protocol; losses for user-side tokens are masked, and supervision is applied only to the assistant's response span containing `[SEG]`. On the image side, we provide a CLIP-preprocessed view for vision–language alignment and a normalized/square-padded view for mask prediction, while preserving both the scaling factors and the original resolution for accurate back-projection.

# 4 Experiment

## 4.1 Experimental Setup

We use the MUSeg underground multimodal semantic segmentation dataset as the primary benchmark. Only the RGB branch participates in training and inference; depth maps are reserved for future extensions. The dataset covers six underground mines of different scales across multiple regions in China, including one training gold mine and five operating mines with diverse geology and commissioning dates. Following a scene/shift split, we adopt 2,247 images for training and 571 for validation.

Training uses a single NVIDIA TITAN RTX (24 GB) GPU. The backbone is LISA-7B, with CLIP ViT-L/14 on the vision side and a SAM-style `mask_decoder` for segmentation. The loss is a weighted sum of CE: 0.5, Dice: 1.5, and BCE: 1.0, and both training and inference use FP16 mixed precision. We employ AdamW with $(\beta_1, \beta_2) = (0.9, 0.95)$, a base learning rate of $3 \times 10^{-5}$, and gradient clipping at 1.0. Constrained by 24 GB of VRAM, we train with DEEPSPEED ZeRO-2 and enable gradient checkpointing to reduce peak memory usage.

We report two metrics: generalized IoU (**gIoU**), the macro-average of per-sample IoU; and class IoU (**cIoU**), computed by accumulating intersections and unions per class and then taking the IoU over foreground classes. We also follow a *no-target* rule: if the ground truth is empty and the prediction is also empty, the sample IoU is set to 1; otherwise, it is set to 0. All results are obtained with batch size = 1 and FP16 inference, ensuring consistency with the training/validation scripts and reproducibility.

## 4.2 Segmentation Results

### 4.2.1 Comparative Experiment

Post-fine-tuning results on the MUSeg dataset indicate that MineLISA significantly outperforms the vanilla baseline in coal-mine segmentation tasks. The generated masks demonstrate robust region coverage and better boundary alignment. Conversely, the unadapted baseline yields negligible results on the test set, validating the importance of our mining-specific adaptation. As shown in Figure 2, MineLISA generates accurate masks for underground targets where the original LISA fails, while retaining general-domain capabilities. The mask quality is quantitatively comparable to human annotation, with precise edge alignment.

### 4.2.2 Ablation Experiment

As shown in Table I, we conducted a joint ablation on loss weights and LoRA rank. With BCE = 1.0 fixed, we compared three settings for (CE, Dice, BCE): $(0.5, 1.5, 1.0)$, $(0.25, 1.75, 1.0)$, and $(0.5, 1.0, 1.0)$. Under the best loss configuration, we further compared LoRA ranks $r = 8$ and $r = 4$.Results show that $(0.5, 1.5, 1.0)$ with $r = 8$ performs best (gIoU = 0.446, cIoU = 0.531), outperforming $(0.25, 1.75, 1.0)$ (gIoU = 0.444, cIoU = 0.526) and $(0.5, 1.0, 1.0)$ (gIoU = 0.430, cIoU = 0.524). Further increasing the Dice weight brings no gain, indicating a trade-off between CE and Dice. With the loss fixed to the optimal setting, $r = 8$ clearly surpasses $r = 4$ (the latter: gIoU =

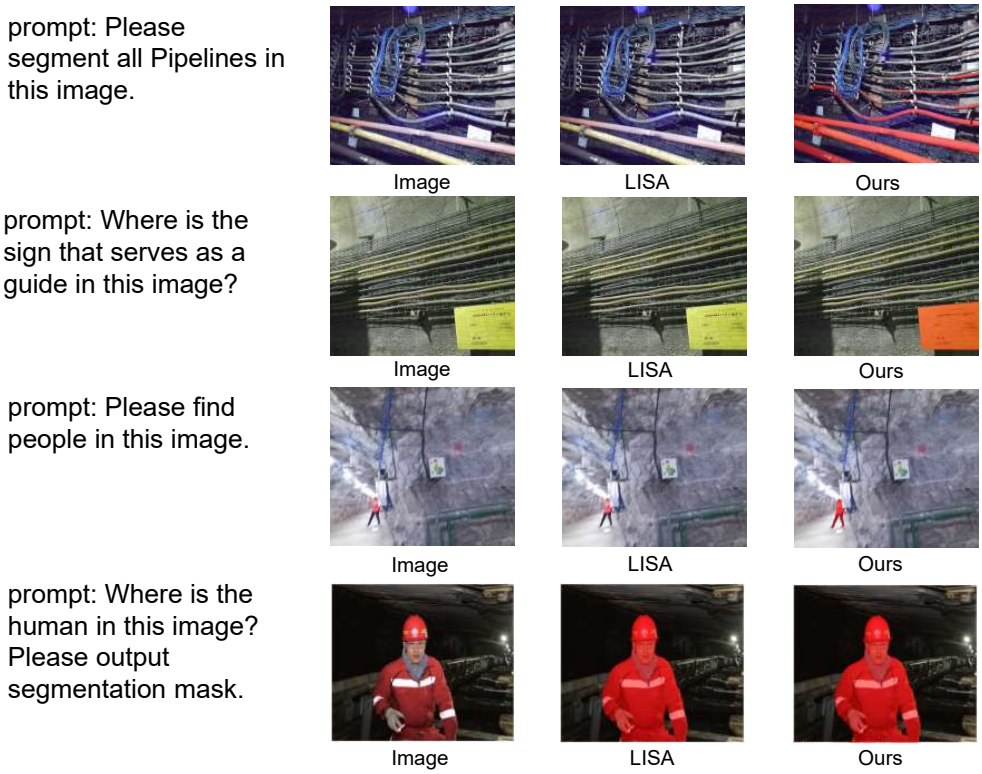


**Fig. 2.** The left panel shows our prompt–response dialogues for different input images. The right panel displays, from left to right, the original image, the segmentation result of the original LISA model, and the result of our model. The red overlays indicate the regions segmented by the models.

0.426, cIoU $= 0.506$), suggesting that a moderate increase in LoRA capacity improves adaptation and segmentation accuracy. Balancing performance and compute cost, we adopt $(\text{CE}, \text{Dice}, \text{BCE}) = (0.5, 1.5, 1.0)$ and $r = 8$ as the default configuration in subsequent experiments.

A key advantage of our approach is its prompt-driven versatility. Operators can leverage natural language to perform on-demand segmentation, switching seamlessly between single-target identification and multi-class parsing. Fig. 3 highlights this instruction-guided flexibility. Unlike traditional static segmentation models, MineLISA leverages the reasoning power of large language models to handle complex queries. Overall, the system provides a robust perception module for complex mining environments, paving the way for autonomous underground operations.

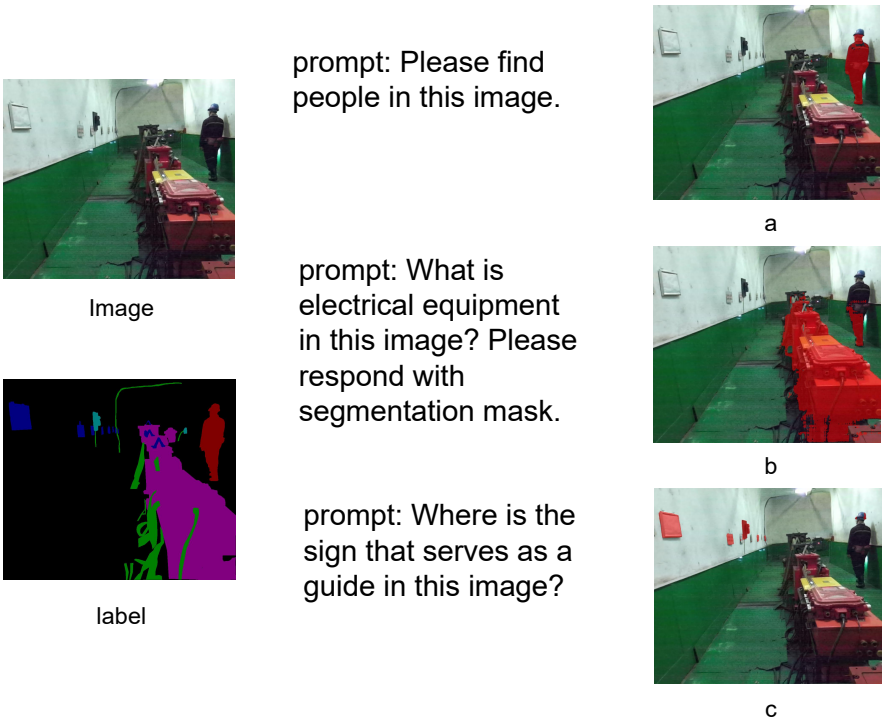


**Fig. 3.** The left panel shows the original image and its multi-class colored labels. The middle column lists the prompts for different categories. The right column (a, b, c) presents the generated results, where red overlays denote the segmented regions produced by the model.

## 5 Discussion and Limitations

Our current approach relies exclusively on RGB imagery. However, underground mines present unique visual challenges, including uneven illumination, airborne dust, and smoke. These factors not only degrade RGB image quality but also impair the accuracy of mask generation; for instance, the model occasionally confuses thick cables with pipes due to texture ambiguity in low-light conditions. Although the MUSeg dataset contains paired Depth information, it remains underutilized in this study.

Depth data possesses the advantage of being invariant to lighting variations and provides critical geometric cues to distinguish foreground equipment from background walls. In future work, we plan to introduce a depth-aware fusion module. By injecting depth embeddings into the vision encoder or fusing them directly at the mask decoder level, we aim to significantly enhance the model's

**Table 1:** Hyperparameter ablation on loss-function weights and LoRA rank: results after 20 epochs of training.

| Loss Weights | | | LoRA | Metrics | |
|---|---|---|---|---|---|
| **CE** | **Dice** | **BCE** | **Rank $r$** | **gIoU** | **cIoU** |
| 0.50 | 1.50 | 1.00 | 8 | 0.446 | 0.531 |
| 0.25 | 1.75 | 1.00 | 8 | 0.444 | 0.526 |
| 0.50 | 1.00 | 1.00 | 8 | 0.430 | 0.524 |
| 0.50 | 1.50 | 1.00 | 4 | 0.426 | 0.506 |

robustness against visual interference in complex mining environments.

## 6 Conclusion

This study applies a multimodal large model to semantic segmentation in coal-mine settings and reports encouraging results under complex underground conditions. With the proposed MineLISA, we move from zero-shot capability to domain proficiency, enabling precise segmentation of multiple target categories in response to natural-language instructions while maintaining robustness to low-light and cluttered environments. This capability allows automatic semantic segmentation to function as an intelligent perception module in hazardous, rapidly changing mines, supporting unmanned operations, real-time decision-making, and improving overall safety. Our experiments show that, with strong semantic understanding and modest domain-specific calibration, large models can effectively generalize to mine-specific scenarios and approach tasks that previously required manual intervention or bespoke architectures. Moreover, the model demonstrates stable performance across diverse object scales and complex background interference. In addition, prompt-driven segmentation offers a new paradigm for human–machine interaction: with a brief instruction, operators can flexibly direct the model to focus on specific risk areas or equipment, substantially improving the efficiency, adaptability, and flexibility of on-site perception.

## Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 52404180, in part by the Fundamental Research Funds for the Central Universities of China under Grant 2024ZKPYZN01, and in part by CHN Energy Science and Technology and Environment Co., Ltd., China "Boundary-aware Intelligent Coal Separation Technology based on Weakly Supervised Training". We thank Associate Professor Ming GU from China University of Mining and Technology-Beijing for language editing.

## Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT in order to: grammar and spelling checking and language polishing. After using this tool, the author(s) reviewed and edited the content as needed and take full responsibility for the publication's content.

## References


[1] Ranjith PG, Zhao J, Ju M, De Silva RV, Rathnaweera TD, Bandara AK. Opportunities and challenges in deep mining: a brief review. Engineering. 2017;3(4):546-51.

[2] Wang G, Ren H, Zhao G, Zhang D, Wen Z, Meng L, et al. Research and practice of intelligent coal mine technology systems in China. International Journal of Coal Science & Technology. 2022;9(1):24.

[3] Shelhamer E, Long J, Darrell T. Fully convolutional networks for semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2017;39(4):640-51.

[4] Lai X, Tian Z, et al.. LISA: Reasoning segmentation via large language model; 2024. ArXiv:2308.00692.

[5] Li S, Kong Q, Gao X, Shi F, Li L, Zhang Q, et al. MUSeg: A multimodal semantic segmentation dataset for complex underground mine scenes. Scientific Data. 2025;12(1):1160.

[6] Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, et al. LoRA: Low-rank adaptation of large language models. ICLR. 2022;1(2):3.

[7] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention; 2015. p. 234-41.

[8] Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL. Semantic image segmentation with deep convolutional nets and fully connected CRFs; 2014. ArXiv:1412.7062.

[9] Chen LC, Papandreou G, Schroff F, Adam H. Rethinking atrous convolution for semantic image segmentation; 2017. ArXiv:1706.05587.

[10] Chen LC, Zhu Y, Papandreou G, Schroff F, Adam H. Encoder-decoder with atrous separable convolution for semantic image segmentation; 2018. ArXiv:1802.02611.

[11] He K, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision; 2017. p. 2961-9.

[12] Xiang Y, Zhao Y, Dong J. Remote sensing image mining area change detection based on improved UNet siamese network. Journal of China Coal Society. 2019;(12).

[13] Jewel MR, Elmahallawy M, Madria S, Frimpong S. Dis-Mine: Instance segmentation for disaster-awareness in poor-light condition in underground mines. In: 2024 IEEE International Conference on Big Data (BigData). IEEE; 2024. p. 6279-88.