Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

JVLGS: Joint Vision–Language Gas Leak Segmentation

Xinlong Zhao
University of British Columbia
Qixiang Pang
University of Central Missouri
Shan Du
shan.du@ubc.ca

University of British Columbia

Research Article

Keywords: Gas leak segmentation, Industrial safety monitoring, Infrared video surveillance system, Spatiotemporal perception, Text-guided feature enhancement, Vision language models (VLMs)

Posted Date: April 21st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9331073/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at The Visual Computer on July 22nd, 2026.
See the published version at https://doi.org/10.1007/s00371-026-04591-y.

JVLGS: Joint Vision–Language Gas Leak Segmentation

Xinlong Zhao[1] · **Qixiang Pang**[2] · **Shan Du**[1]

Received: date / Accepted: date

Abstract Gas leaks pose severe risks to human health and industrial safety. However, accurate and timely monitoring of gas leaks remains a major challenge. Existing vision-based methods using infrared (IR) imagery are limited by the inherently blurry and non-rigid nature of leak plumes, which reduces detection reliability and precision. To overcome these limitations, this paper proposes a Joint Vision–Language Gas leak Segmentation (JVLGS) framework that integrates the complementary strengths of visual and textual modalities to enhance gas leak segmentation. Recognizing that gas leaks are sporadic and many video frames contain no leakage, JVLGS incorporates an adaptive postprocessing module to effectively suppress false positives caused by noise and non-target objects—a common limitation of existing approaches. Extensive experiments across diverse industrial scenarios demonstrate that JVLGS significantly outperforms state-of-the-art gas leak segmentation methods. Furthermore, it achieves consistently strong performance under both supervised and few-shot learning settings, whereas competing methods typically perform well in only one setting or underperform in both. We publish our code at: `https://github.com/GeekEagle/JVLGS`.

Keywords Gas leak segmentation · Industrial safety monitoring · Infrared video surveillance system · Spatio-temporal perception · Text-guided feature enhancement · Vision language models (VLMs).

Shan Du *
shan.du@ubc.ca

Xinlong Zhao
zhaoxinlong1@gmail.com

Qixiang Pang
qpang@ucmo.edu

[1]Irving K. Barber Faculty of Science, University of British Columbia Okanagan, Kelowna, BC V1V 1V7, Canada

[2]Department of Computer Science and Cybersecurity, University of Central Missouri, Warrensburg, MO 64093, USA

1 Introduction

Industrial gas leaks constitute critical emergencies that can lead to severe accidents. Methane contributes significantly to the greenhouse effect, while others like acetaldehyde are highly toxic [1]. Detecting these leaks is especially challenging because most industrial gases are colorless and odorless, making them imperceptible to the human senses. Traditional detection methods rely on electrochemical or photoionization sensors to monitor pipeline conditions [38]. These approaches are hazardous because they require inspectors to be close to suspected leak points. To mitigate these hazards, vision-based detection methods have gained increasing attention due to their non-contact nature and potential for automation [36]. Such methods reduce labor costs while eliminating direct exposure risks for inspectors [25]. Among them, video surveillance integrated with infrared (IR) imaging has shown strong potential for detecting gas leaks.

Nevertheless, video surveillance systems still face significant challenges. Gas leaks are difficult to detect, even with infrared cameras, because they blend seamlessly into the background. To address this issue, Lu et al. [17] proposed a Gaussian-based background modeling method for adaptive gas leak segmentation. However, its effect degrades significantly in the presence of irregular gas shapes and dynamic backgrounds. Similarly, optical flow-based methods [34] are limited by inconsistent gas shapes and are susceptible to noise interference.

Recent advances in deep learning have led to significant progress in camouflaged object detection (COD) tasks [30]. For example, SINet-V2 [5] outlined the rough contours of concealed objects, and GasSeg [37] combined con-

volutional neural networks (CNNs) with Vision Transformers (ViT) to develop a lightweight model for gas leak segmentation. However, they struggle to capture motion cues from faint and ambiguous targets. Long-term motion-based techniques [2,3] face difficulties due to the highly variable shapes of gas leaks and the lack of consistent features, while short-term methods, such as RCRNet [33], leverage optical flow but remain susceptible to noise. Implicit motion analysis methods, like SLT-Net [4], show potential but struggle to distinguish gas leaks from texturally similar objects, such as clouds. FGSTP [41] introduced fine-grained perception to refine motion information; however, the extracted features are not always reliably correlated with actual targets.

One promising direction to enhance feature representation is the integration of temporal-spatial models with Vision-Language Models (VLMs) [39]. For instance, CLIP [23] uses a ViT as both image and text encoder, but is too slow for real-time segmentation. VisualBERT [14] enables end-to-end vision-language correlation learning but lacks spatial precision. DALL-E [24] combines visual information with large language models (LLMs) for creative generation but is unsuitable for precise segmentation. OWLv2 [18] employs a ViT to process both image and text inputs via self-training and shows potential for open-vocabulary detection. Nevertheless, few existing methods effectively incorporate text-guided segmentation for blurry, non-rigid objects such as gas leaks.

In this paper, we propose Joint Vision-Language Gas leak Segmentation (JVLGS), a novel framework designed to enhance model generalizability across diverse gas leak scenarios. Our method integrates text prompts between the vision encoder and motion extractor to enrich feature representations. Following vision–language integration, a temporal-spatial module is designed to analyzes motion and spatial cues of gas leaks using these enhanced features. The decoder then integrates multi-scale information through upsampling to generate the final masks. Many prior methods [21,4] are vulnerable to noise and minor movements in non-leak frames, often producing false positives when no target is present. To address this, JVLGS incorporates a post-processing step that eliminates false positives, ensuring completely blank masks for non-leak cases. We evaluate our model on two benchmark datasets under both supervised and few-shot learning settings to simulate realistic video surveillance scenarios.

The main contributions of this work are summarized as follows:

- We propose JVLGS, a framework that integrates manually crafted text prompts with a video clip to enhance feature extraction. These prompts enable accurate localization and detection of gas leaks.
- We design a lightweight vision-language fusion approach, enabling more effective integration of visual features with textual representations.
- JVLGS achieves significant performance gains in both supervised and few-shot learning, effectively simulating real-world leak detection scenarios.
- Among all benchmark models, JVLGS delivers a favorable trade-off between segmentation accuracy and computational efficiency, making it suitable for practical industrial monitoring applications.

2 Related Work

2.1 Gas Leak Detection and Segmentation

Traditional gas leak detection and segmentation methods rely on manual sampling and laboratory analysis based on chemical or stoichiometric reactions. Although these approaches are generally accurate, they suffer from slow response times, high labor costs, and safety risks associated with manual gas sampling [42].

Recent advancements in infrared (IR) thermography have made vision-based detection a promising alternative [38]. Vision-based methods offer convenient operation, wide detection coverage, real-time feedback, and intuitive visualization of gas leaks. However, automatically segmenting invisible gas remains challenging. Existing vision-based methods can be broadly classified into image-based and video-based approaches. Image-based methods [31,35] depend on local visual features, such as shape and texture. However, their performance degrades when dealing with blurry or low-contrast gas regions in infrared imagery. In contrast, video-based methods [27,12] exploit the motion characteristics of gas in video sequences, which typically improves detection accuracy. Nevertheless, these approaches often incur high computational costs and are unsuitable for real-time industrial applications. To overcome these limitations, we propose a video-based framework that effectively integrates visual and textual information to enhance gas leak segmentation. The proposed model achieves state-of-the-art performance while maintaining practical efficiency for industrial deployment.

2.2 Camouflaged Object Detection

Camouflaged Object Detection (COD) aims to identify objects that blend seamlessly into their surroundings, often exhibiting low contrast and ambiguous boundaries. Traditional COD methods [6] rely on hand-crafted features such as texture, color, or edge cues. However, these approaches perform poorly in complex environments or low-resolution imagery due to their limited representational capacity. With the advent of deep learning, bioinspired approaches have been

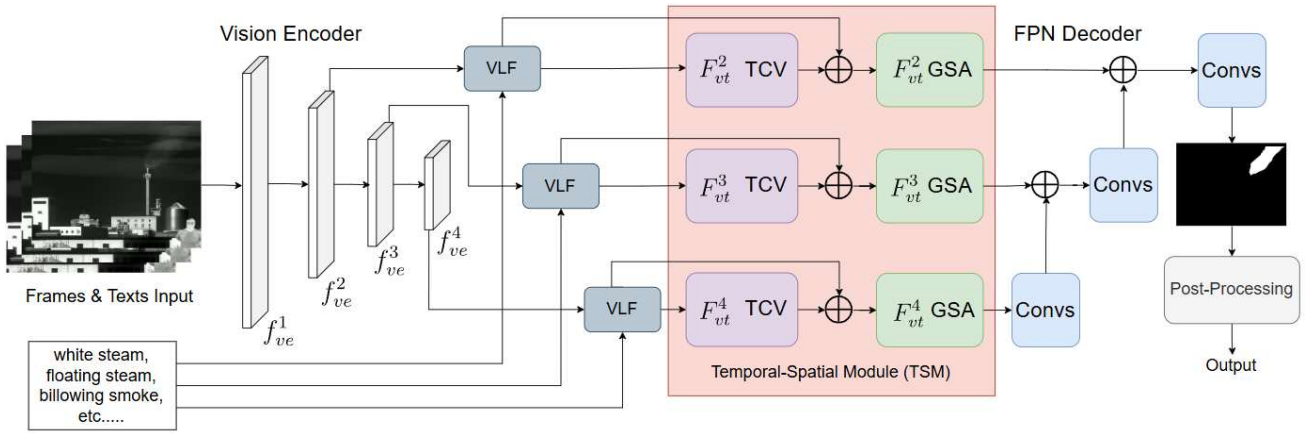


Fig. 1: Framework of Joint Vision-Language Gas leak Segmentation (JVLGS). The inputs are video clips and text prompts of the target object. Visual and textual features are fused in the VLF module, followed by the Temporal-Spatial Module (TSM) and decoder. The post-processing removes false positives for final masks.

introduced to enhance detection efficiency and robustness [32]. Although these methods achieve real-time inference, they remain less effective for detecting small or blurry objects. Motion-based COD approaches further exploit temporal information by incorporating cross-frame cues and motion coherence modules to improve segmentation precision [8,13]. To overcome the limitations of two-stage motion-based pipelines, recent end-to-end models perform object localization and segmentation within a unified framework, thereby suppressing distractions from background motion [10,40]. These models leverage temporal continuity to ensure inter-frame consistency of detected objects.

Inspired by these advancements, this paper proposes an end-to-end framework specifically designed for gas leak monitoring in complex and unstable industrial environments. The proposed approach effectively captures subtle, motion dependent leak patterns that closely resemble camouflaged objects in dynamic infrared scenes.

2.3 Vision-Language Model

Vision–Language Models (VLMs) map both visual and textual modalities into a shared embedding space, thereby enhancing the visual features [39]. Among early foundational models, CLIP [23] employs a dual-encoder architecture to extract visual and textual features and measure their similarity through contrastive learning. However, obtaining high-quality image-text pairs is time-consuming and resource-intensive. Region-based method, MDETR [11], integrates contrastive pretraining with region-level supervision to achieve precise localization. However, these methods rely heavily on well-defined boundaries, making them less robust in ambiguous or indistinct scenarios, such as gas leaks. GLIP [15] combines contrastive learning with phrase grounding, detecting objects first and then refining via text-guided attention; yet they risk overfitting and limit generalization to diffuse leak patterns. More recently, OWLv2 [18] has advanced text-guided object detection through self-training with pseudo-annotations, extending the prompt-based framework of OWL-ViT [19]. Its pretrained text encoder demonstrates strong alignment between textual prompts and visually ambiguous objects, making it particularly well-suited for applications involving subtle or non-distinct visual cues. In this paper, we leverage the pretrained text model of OWLv2 as our text prompt module to construct a state-of-the-art gas leak segmentation model.

3 Method

3.1 Model Architecture

Fig. 2 illustrates the overall architecture of the proposed Joint Vision–Language Gas leak Segmentation (JVLGS) framework. The model comprises the following components: (1) Vision encoders based on Pyramid Vision Transformer-v2 (PVTv2) [28]; (2) A Vision-Language Fusion (VLF) module designed to integrate visual features with text representations effectively; (3) A Temporal-Spatial Module (TSM) proposed to capture fine-grained motion and spatial information related to target objects; (4) An enhanced decoder built upon the Feature Pyramid Network (FPN) [16], optimized for multi-scale feature decoding; (5) A post-processing step to remove false positives. Overall, JVLGS takes video clips and text prompts as inputs and generates a binary mask for each frame, indicating the presence or absence of gas leakage.

Numerous studies have demonstrated that text prompts can enrich semantic representations, enabling models to bet-

ter distinguish target objects in complex visual scenes [20, 9]. Our proposed JVLGS integrates textual representations into visual features, enabling the Temporal-Spatial Module (TSM) to localize targets more accurately through text-guided encoding. Furthermore, we fine-tune the pretrained text model of OWLv2 [18] as our text module.

3.2 Vision-Language Fusion

The input video clips are first processed and divided by the vision encoder, PVTv2 [28] into three scale features $f_{ve}^{i} \in R^{C_v \times H/2^{i+1} \times W/2^{i+1}}, i \in \{2,3,4\}$, where C_v, H, W are channels, height and width of a frame. Fig. 2 shows the details of the Vision-Language Fusion (VLF) module. The input of VLF, V_e, consists of $f_{ve}^{i}, i \in \{2,3,4\}$. The size of V_e is $T \times C_v \times H/2^{i+1} \times W/2^{i+1}, i \in \{2,3,4\}$, where T is the length of the video clip. The VLF firstly permute V_e to $V_e^T \in R^{T \times H/2^{i+1} \times W/2^{i+1} \times C_v}$ $\{2,3,4\}$. Then V_e^T is linearly transformed to the vision feature f_v.

The text encoder in JVLGS is adapted from the OWLv2 text module, which employs a Vision Transformer (ViT) backbone and was originally developed for zero-shot object detection [18]. Although OWLv2 outputs bounding boxes, in our framework, its textual embeddings are repurposed to guide visual attention toward target regions, thereby enabling the generation of more precise binary segmentation masks. The VLF module is designed to integrate visual cues with textual guidance, where the vision features support the localization of target objects, and the text representations enrich visual features. Considering the high variability of industrial video content and the linguistic limitations of pretrained language models, multiple descriptive prompt words are utilized to capture diverse gas leak patterns. Since gas leaks are typically invisible to the human eye and standard cameras, many pretrained vision–language models lack a direct understanding of the concept of a "gas leak". However, we observed that the synonyms of leaks, visually, such as billowing smoke or white steam, are effective for the language module, because leaks appear similar to these in infrared video. Thus, we introduce these words as our prompt words, and we use a total of four prompt sequences during training. The validation of our selective prompts can be seen in Section *IV.D*. We denote the text input as $W \in R^{P_l \times d}$, where P_l is the number of the prompt sequence and d is the dimension of each prompt sequence. Then the prompted words are encoded by the text encoder as follows:

$$k, A = TextEncoder(W) \tag{1}$$

where k is the key of the prompt sequences, and A is an attention mask, indicating which tokens are correlated inputs for the text model. The k and A are then fed into the text

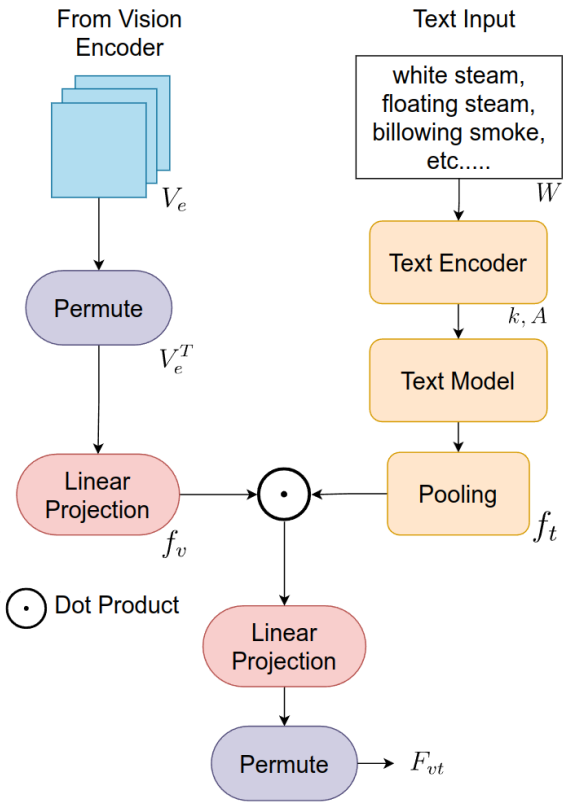


Fig. 2: Architecture of the Vision-Language Fusion (VLF) module. The vision encoder extracts visual features, while text prompts are generated from keywords or synonyms of the target objects.

encoder to obtain the textual representations as follows:

$$f_t = TP(OwlText(k, A)) \tag{2}$$

where $OwlText(k, A)$ represents the extracted text representations $f_t \in R^{P_l \times C_t}$, and C_t is the channel size of text representation. $TP(\cdot)$ denotes token-level pooling that selects a token from the hidden state of each sequence to refine f_t. Next, the text representation fuses with the f_v. The fusion operation in each scale is as follows:

$$F_{vt} = (\phi(f_v \cdot f_t))^T \tag{3}$$

where $\phi(\cdot)$ is the linear projection that converts the channel number of representative text C_t, to the channel number of vision features, C_v.

The resulting vision-language features, F_{vt}, are obtained by fusing the outputs of the vision and text encoders. After applying a linear projection to align feature dimensions, we fuse them through a dot product and then reshape the result to match the output shape of the vision encoder. The final fused feature, F_{vt}, is subsequently fed into the Temporal-Spatial Module (TSM) to capture motion and spatial information.

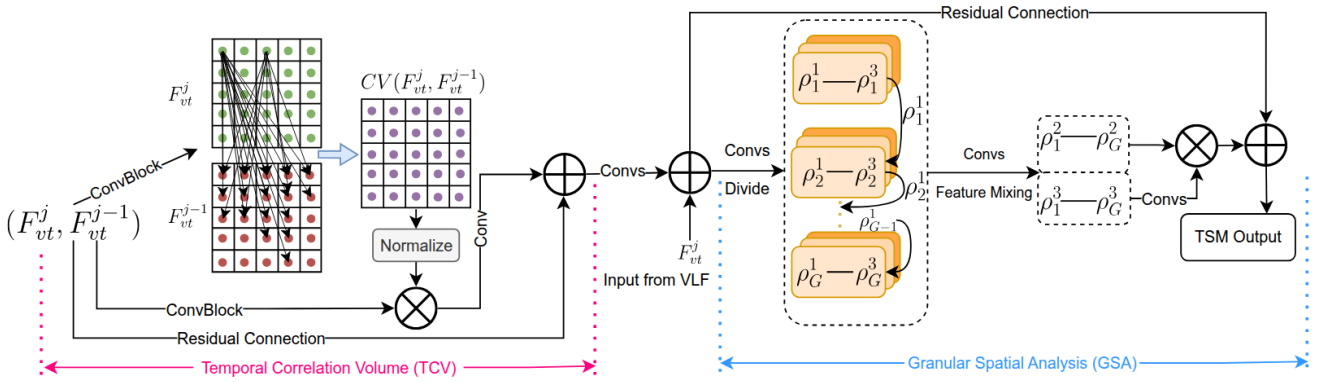


Fig. 3: Framework of Temporal-Spatial Module (TSM), which is constituted by two parts, Temporal Correlation Volume (TCV) and Granular Spatial Analysis (GSA). The former is to find the target object through inter-frame changing, while the latter refines the intra-frame object features based on the TCV output.

3.3 Temporal-Spatial Module

The Temporal-Spatial Module (TSM) is proposed to capture implicit motion information of the target object and refine it using the text-enhanced features produced by the Vision-Language Fusion (VLF) module. The architecture of TSM is illustrated in Fig. 3. It consists of two key components: the Temporal Correlation Volume (TCV) and Granular Spatial Analysis (GSA).

The TCV is to obtain strong motion features through the temporal correspondence, based on the combination of vision-text features in frames j and $j-1$, $(F_{vt}^{j}, F_{vt}^{j-1}) \in R^{2C\times H'\times W}$ where C is the channel number, H' and W' are height and width of the feature map. Inspired by [4], the TCV computes the correlation volumes between two feature maps of consecutive frames F_{vt}^{j} and F_{vt}^{j-1} to find subtle motion cues. The 4D correlation volume $CV(F_{vt}^{j}, F_{vt}^{j-1}) \in R^{H'\times W'\times H'\times W'}$, is as follows:

$$CV(F_{vt}^{j}, F_{vt}^{j-1})_{xyuv} = \exp\left(\sum_{c}(F_{vt}^{j})_{xyc}\cdot(F_{vt}^{j-1})_{uvc}\right) \tag{4}$$

where c denotes the channel index, and $xyuv$ are the coordinates of the current and adjacent frames. The $CV(F_{vt}^{j}, F_{vt}^{j-1})_{xyuv}$ denotes the correlation volumes between two elements at positions (x,y) and (u,v) in feature maps F_{vt}^{j} and F_{vt}^{j-1}. Inconsistencies in the same positions (x,y) between two feature maps indicate motion objects nearby. The Temporal Correlation Volume (TCV) module subsequently determines whether the motion in this area is relevant to the target object to distinguish and track. For each feature map F_{vt}^{j} of a frame j, the TCV computes two correlation volumes between current frame and two adjacent frames, $CV(F_{vt}^{j}, F_{vt}^{j-1})$ and $CV(F_{vt}^{j}, F_{vt}^{j+1})$, to strengthen the reliability of the extracted features.

Normalization is then applied to the last two dimensions of $CV(f_t, f_{t-1})_{xyuv}$, (u,v), to refine the relevant motion features between reference frames and the current frame. The normalization is formulated as follows:

$$\widetilde{CV}(F_{vt}^{j}, F_{vt}^{j-1})_{xyuv} = \frac{CV(F_{vt}^{j}, F_{vt}^{j-1})_{xyuv}}{\sum_u\sum_v CV(F_{vt}^{j}, F_{vt}^{j-1})_{xyuv}} \tag{5}$$

Next, the normalized correlation volumes are aggregated with the channel-wise feature map $\lambda(F_{vt}^{j}, F_{vt}^{j-1}) \in R^{C\times H'\times W'}$, where $\lambda(\cdot)$ is a convolution layer that converts the channel number of $(F_{vt}^{j}, F_{vt}^{j-1})$, $2C$ to C. At last, the TCV obtains the final correlation volume $\varphi(F_{vt}^{j-1,j})$. The calculation of another correlation volume $\varphi(F_{vt}^{j+1,j})$ is the same as $\varphi(F_{vt}^{j,j-1})$. The aggregation of the channel-wise features and normalized correlation volumes is as follows:

$$\varphi(F_{vt}^{j,j-1}) = \lambda(F_{vt}^{j}, F_{vt}^{j-1})\cdot\widetilde{CV}(F_{vt}^{j}, F_{vt}^{j-1})_{xyuv} \tag{6}$$

While the TCV captures motion information, it cannot refine this into detailed spatial representations within each frame. To address this limitation, we propose a Granular Spatial Analysis (GSA) module to convert the extracted correlative motion into specific object features, inspired by [21]. The input of GSA $\tau(F_{vt}^{j,j-1})$ combines $(\varphi(F_{vt}^{j,j-1})$ and F_{vt}^{j} from VLF as follows:

$$\tau(F_{vt}^{j,j-1}) = \varphi(F_{vt}^{j,j-1}) + F_{vt}^{j} \tag{7}$$

First, the 1×1 convolution expands the channel dimension of $\tau(F_{vt}^{j,j-1})$. Then the module divides the expanded $\tau(F_{vt}^{j,j-1})$ into g groups $(G_1 - G_g)$ according to channels, and each group includes three channels of features $\rho_i^1, \rho_i^2, \rho_i^3$. Next, we split the first channel ρ_i^1 of G_i to concatenate it to the last two layers of the next group of features $(\rho_{i+1}^2, \rho_{i+1}^3)$. The overall processes are as follows:

$$\begin{aligned} \rho_i^1 &= Split(Expand(G_i)) \\ G_{i+1} &= Concat(G_{i+1}, \rho_i^1) \\ \rho_{i+1}^1 &= Split(Expand(G_{i+1})) \end{aligned} \tag{8}$$

The mixing is designed to analyze granular spatial cues across different channel groups, resulting in a more robust and informative feature representation. Functionally, the iterative operation is equivalent to integrating a multi-path pyramid structure. Next, the second and third feature groups ($\rho_1^2 - \rho_G^2$ and $\rho_1^3 - \rho_G^3$) are merged into two new groups, N_2 and N_3, respectively, through concatenation and convolution. Finally, the two sets of group-mixed features (N_2 and N_3) are fused through a convolutional block with a residual connection. The output of the GSA is denoted as gsa_{vt}^j.

$$\begin{aligned} N &= N_2 \cdot N_3 \\ gsa_{vt}^j &= \tau(F_{vt}^{j,j-1}) + Conv(N) \end{aligned} \tag{9}$$

The TSM module effectively extracts valuable temporal and spatial cues across video frames and feature channels, enabling rich information interaction and progressive feature refinement. This module strengthens the model's ability to perceive the target objects and provides a solid foundation for the final decoder predictions.

3.4 Decoder and Post-Processing

The decoder is a modified FPN that integrates multi-scale information by upsampling. Leveraging the rich features produced by the preceding modules, the FPN alone is sufficient to generate accurate segmentation masks. In contrast, using a separate COD head, like [5], may degrade the stability of the features. The effectiveness of our decoder design is further demonstrated in Section *IV.E*.

Unlike typical VOS studies, which assume that target objects are present in every frame of a video [2,9], gas leaks do not appear sporadically throughout a video. Thus, the expected output of those non-leak frames should be completely black masks. However, significant noise results in false-positive masks for non-leak frames and makes many models fail in producing clean, completely black masks. To address the problem, we employ morphological opening [26] as a post-processing. Opening is defined as the dilation of the erosion of a binary mask by a kernel. The effectiveness of the opening depends on the kernel size. An oversized kernel may remove true-positive regions during the erosion stage, which cannot be restored during dilation, while an undersized kernel may fail to eliminate false positives. In our framework, the kernel size is adjustable according to different scenarios. Overall, this post-processing effectively suppresses false positives while preserving the integrity of true leak regions.

3.5 Loss Function

To facilitate comparison with benchmark models, we adopt the commonly used loss function in the field of video object segmentation, Binary Cross-Entropy (BCE) and Intersection over Union (IoU). The BCE emphasizes pixel-level accuracy while the IoU captures global structure similarity. However, both loss functions treat all pixels equally and do not account for the varying importance of different regions. A more effective approach is to assign pixel-wise weights to specific areas, such as object boundaries and centers, which are typically more challenging to segment. If the value of pixel (i, j) is different from its surrounding pixels, it indicates the place is more important. Thus, the place in this frame will be largely weighted. The formulas for weighted BCE L_{bce}^w and weighted IoU L_{iou}^w are the following:

$$\begin{aligned} L_{\text{bce}}^w &= \frac{\sum_{k=1}^{HW}(1+\alpha w_k)\left[y_k \log p_k + (1-y_k)\log(1-p_k)\right]}{\sum_{k=1}^{HW}(1+\alpha w_k)} \\ L_{\text{iou}}^w &= 1 - \frac{\sum_{k=1}^{HW}(1+\alpha w_k)\left(y_k p_k\right)}{\sum_{k=1}^{HW}(1+\alpha w_k)\left(y_k + p_k - y_k p_k\right)}. \end{aligned} \tag{10}$$

where α is the hyperparameter, p_{ij} and gt_{ij} are the prediction probability and the ground truth pixel at coordinate (i, j). w_{ij} is the weight for each pixel. Inspired by [29], we add weights to each position of a ground truth frame in both loss functions as follows:

$$w_{ij} = \left| \frac{\sum_{S_{ij}} gt}{\sum_{S_{ij}} 1} - gt_{ij} \right| \tag{11}$$

where S_{ij} is the surrounding area of the pixel (i, j). In a word, weighted BCE is assigned weights for pixels to focus on critical areas. The weighted IoU loss is a crucial component for video object segmentation. It highlights challenging pixels while retaining attention to the overall structural accuracy of the object masks. Finally, our loss function combines the wBCE and wIoU, defined as follows:

$$L = L_{bce}^w + L_{iou}^w \tag{12}$$

4 Experiments

4.1 Datasets

We evaluate our model on two datasets. The first, SimGas, is a synthetic dataset. Its backgrounds are partially sourced from a real-world dataset, GasVid [27], while the rest are generated using DALL·E 2 [24]. Foreground gas leaks are simulated using Blender. Manual annotation of GasVid frames is challenging due to the difficulty of precisely delineating leak regions, and the limited variety of leak scenarios constrains the evaluation of model generalizability. In comparison, the SimGas dataset provides more diverse leak scenarios that better represent real-world conditions. Notably,

Table 1: Quantitative comparison with benchmark methods on two datasets.

Models	Input Type	Year	SimGas (Avg)			IGS-Few		
			$\mathscr{J}\uparrow$	$\mathscr{F}\uparrow$	$\mathscr{J}\&\mathscr{F}\uparrow$	$\mathscr{J}\uparrow$	$\mathscr{F}\uparrow$	$\mathscr{J}\&\mathscr{F}\uparrow$
SINet-V2 [5]	Image	2021	37.29	42.76	40.03	17.45	28.15	22.80
GasSeg [37]	Image	2024	37.35	46.16	41.76	55.40	67.00	61.20
SLT-Net [4]	Video	2022	39.24	44.40	41.82	66.43	76.52	71.47
Cutie [3]	Video	2024	-	-	-	46.61	59.07	52.84
Zoomnext [21]	Video	2024	21.67	27.10	24.39	59.28	70.43	64.86
LangGas [7]	Video	2025	53.57	63.95	58.76	10.66	15.16	12.91
FGSTP [41]	Video	2025	39.45	44.42	41.94	66.70	76.56	71.63
JVLGS (Ours)	Video	2025	**61.73**	**70.25**	**65.99**	**67.06**	**77.05**	**72.05**

it preserves precise annotations by compositing realistically rendered leaks and occluding foreground elements onto backgrounds. The dataset comprises 31 videos, totaling approximately 1,600 frames. Additional details are provided in [7].

The second dataset is derived from the IGS dataset [37], which contains 4,453 video frames for training and 1,897 for testing across a total of 27 videos. In the original dataset split, training and testing frames are drawn from the same videos, with the majority of frames from each video used for training and the remaining frames reserved for testing. However, video surveillance systems always adopt few-shot training, which uses only a few frames per video for training, while the majority are reserved for testing. To better reflect real-world leak monitoring conditions, we adopt this few-shot training strategy. Specifically, we cap the number of training frames per video, leaving hundreds of frames per video for testing. We refer to this reconfigured dataset as IGS-Few.

4.2 Implementation Details

For fair comparisons across methods, all experiments are conducted using the same software and hardware configurations. We set the experimental image resolution to 352×352. We train for 60 epochs on SimGas and 150 epochs on the IGS-Few dataset. Moreover, we set the initial learning rate (lr) to 10^{-4}, which decays to 10^{-5} in the final 20% of training epochs. Following standard practices in VOS evaluation, we adopt metrics that assess both region similarity and boundary accuracy between the predicted masks and the ground truth annotations. Empirically, we adopt the Jaccard Index ($\mathscr{J}$), contour accuracy ($\mathscr{F}$), and their average ($\mathscr{J}\&\mathscr{F}$), following [22], to evaluate mask overlap and boundary alignment. All experiments are performed on an NVIDIA RTX 4090 GPU with 24 GB of memory.

4.3 Results

4.3.1 Experimental Setup

Different models exhibit significant variations in both evaluation metrics and dataset splits. LangGas [7] computes $\mathscr{J}$ and $\mathscr{F}$ for each video by confusion matrices, while others calculate the average $\mathscr{J}$ and $\mathscr{F}$ in all frames of the test set. GasSeg [37] uses most frames for training and minors for testing. To ensure a fair comparison, we standardize metric computation for all benchmark models in *IV.C*.2. Additionally, we report results using each method's original metric computation or dataset split in *IV.C*.3.

4.3.2 Comparison with SOTA Methods

Table 1 summarizes the results using our unified metric computation. And Fig. 4 and Fig. 5 present qualitative visualizations on the SimGas and IGS-Few datasets. SINet-V2 is an early method for camouflaged object detection and does not have a temporal module. Consequently, its performance is lower than many other methods on both datasets. GasSeg effectively finds leaks by fusing local and global information. It performs well on its proposed dataset, but it struggles on SimGas, resulting in a low $\mathscr{J}\&\mathscr{F}$ of 41.76. SLT-Net and FGSTP demonstrate comparable performance on both the SimGas and IGS-Few datasets. However, they suffer from false positives in non-leak frames, often generating random artifacts. ZoomNeXt focuses on capturing multi-scale spatial features but overlooks motion information. Thus, it underperforms on the SimGas videos. Cutie depends on the mask of the first frame to initialize the target object and guide segmentation in subsequent frames. The absence of leaks in the initial frames poses a huge obstruction to the SimGas dataset prediction. LangGas combines background subtraction with a large language model for mask prediction. But it struggles with the dithering backgrounds. Our pro-

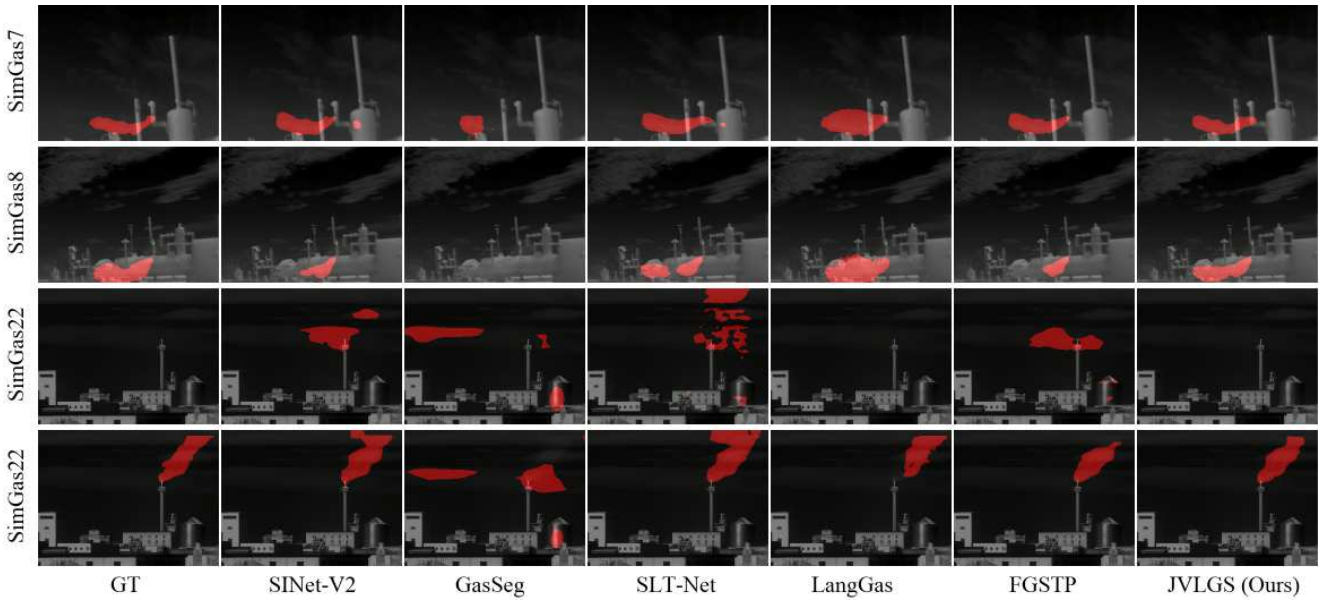


Fig. 4: Visualized results on SimGas. Our model achieves the highest accuracy and generates whole-black masks for non-leak cases.

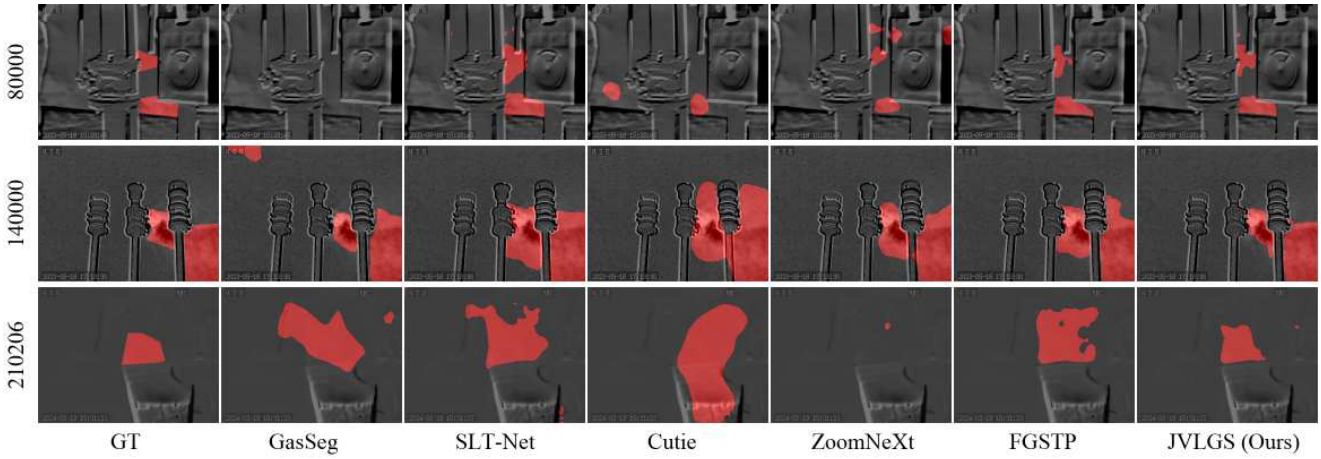


Fig. 5: Visualized results on IGS-Few. Our model distinguishes the leak better than other models in few-shot learning.

posed method, JVLGS, integrates motion, spatial, and textual prompts to enhance feature representation. It achieves a $\mathcal{J}\&\mathcal{F}$ score of 65.99 on SimGas and 72.05 on IGS-Few, outperforming all other methods on both datasets.

4.3.3 Different Metrics Comparison with Some Models

As noted before, LangGas and GasSeg use different evaluation metrics and dataset splits from ours. To ensure fairness, we re-implement their calculations to evaluate our method and LangGas. Table 2 presents results based on LangGas's metric computation, while Table 3 shows performance under GasSeg's dataset split. In both settings, our method consistently outperforms the others, demonstrating its robustness across evaluation protocols.

Table 2: Comparison with LangGas using its original metric calculation, on the SimGas dataset.

Models	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$	$\mathcal{J}\&\mathcal{F}\uparrow$
LangGas [7]	58.82	72.82	65.82
JVLGS (Ours)	**69.54**	**79.75**	**74.65**

Table 3: Comparison with GasSeg using its original split, on the IGS dataset.

Models	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$	$\mathcal{J}\&\mathcal{F}\uparrow$
GasSeg [37]	78.21	86.42	82.31
JVLGS (Ours)	**81.82**	**89.01**	**85.42**

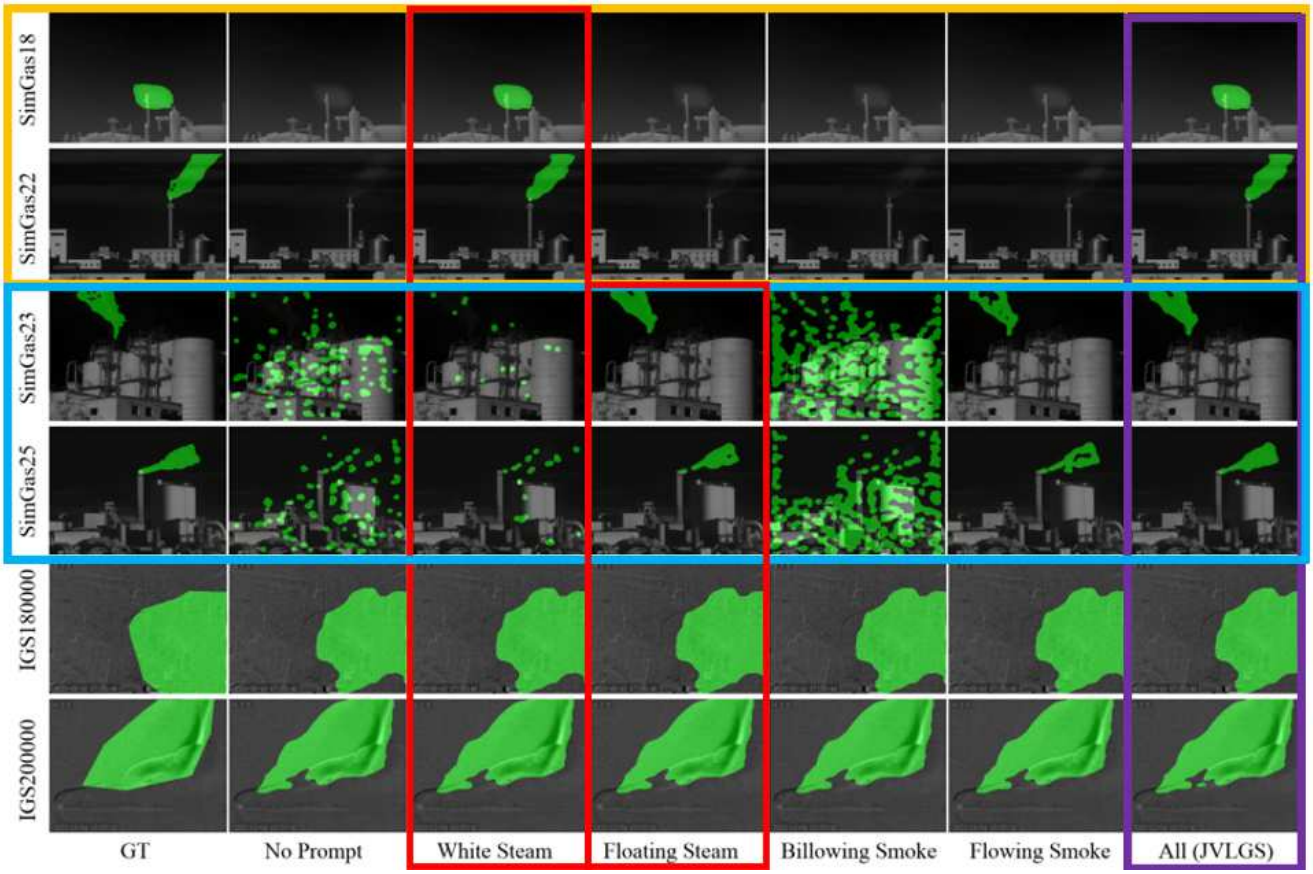


Fig. 6: Visualized contribution of text prompt. Each prompt word tends to enhance performance in some specific scenarios. e.g, "White Steam" contributes significantly to videos 18 and 22. "Floating Steam" contributes to videos 23 and 25.

4.4 Analysis of Text Prompt Effect

We further verify the effectiveness of the language prompt. Specifically, we calculate the improvements of $\mathcal{J}\&\mathcal{F}$ by different prompts for each video on both datasets. Table 4 quantifies the contribution of each prompt on the SimGas dataset. We define prompt contribution as the average increase in $\mathcal{J}\&\mathcal{F}$ score relative to the non-prompt baseline, calculated across all videos. The column "Improved Videos" indicates the number of videos significantly influenced by each prompt. A video is considered "improved" if the improvement in its $\mathcal{J}\&\mathcal{F}$ score exceeds 0.5 points.

The impact of each prompt is also visually illustrated in Fig. 6. The video frames located at the intersection of red, yellow, and blue rims indicate that different prompts benefit specific subsets of videos. Missing or incorrect prompts are ineffective at localizing leaks. This observation underscores the importance of using multiple prompts during training. Ultimately, all four prompt words collectively enhance feature representations on the SimGas dataset, ensuring that each gas leak video is guided by at least one effective prompt. However, the last two rows of the figure indicate that the prompts have only a slight influence on predictions for the IGS-Few dataset. Thus, the language prompts are primarily beneficial for unseen scenarios. In contrast, the proposed model is capable of achieving reliable segmentation for seen scenarios even under few-shot training without the need for additional language prompts.

Table 4: Contribution of each prompt for JVLGS performance on the SimGas dataset.

Prompts	Avg. $\mathcal{J}\&\mathcal{F}$ Contribution	No. of Improved Videos
No Prompt	0	0
"White Steam"	15.03	21
"Floating Steam"	10.26	12
"Billowing Smoke"	1.14	12
"Blowing Smoke"	8.17	18
All (JVLGS)	26.35	28

4.5 Ablation Studies

Table 5 presents the results of our ablation studies conducted on both the SimGas and IGS-Few datasets. We evaluate the

Table 5: Ablation studies on both datasets demonstrating the effectiveness of each component within our proposed model.

TSM	VLF	Decoder	Post-Processing	SimGas (Avg)			IGS-Few		
				$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$	$\mathcal{J}\&\mathcal{F}\uparrow$	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$	$\mathcal{J}\&\mathcal{F}\uparrow$
✓		NCD		39.45	44.42	41.94	66.70	76.56	71.63
✓	✓	NCD		38.88	44.24	41.55	66.90	76.96	71.93
✓	✓	NCD	✓	57.73	66.05	61.90	66.94	76.97	71.95
✓		FPN	✓	59.21	67.26	63.23	66.97	**77.08**	72.03
✓	✓	FPN		39.06	44.50	41.78	67.02	77.05	72.03
✓	✓	FPN	✓	**61.73**	**70.25**	**65.99**	**67.06**	77.05	**72.05**

individual contributions of the VLF module, FPN decoder, and post-processing.

Effectiveness of Text Prompts. We assess the impact of the language prompt module while keeping all other components unchanged. As shown in Rows 4 and 6 of Table 5, removing the VLF module results in performance degradation on both datasets, with a more noticeable drop on SimGas and a slight decrease on IGS-Few. This observation highlights the greater contribution of the VLF module on the SimGas dataset compared to IGS-Few.

Decoder Choice. The differences between Rows 2 and 5, as well as Rows 3 and 6 in Table 5, underscore the importance of decoder choice. Under identical conditions, regardless of the presence of the VLF module and post-processing, models using the FPN decoder consistently outperform those using the NCD decoder [41] on both datasets. This can be attributed to FPN's straightforward multi-scale feature integration. In contrast, the NCD involves more complex structures, such as reversing and inserting feature maps, which may lead to incompatibilities with other modules and ultimately hinder overall performance.

Effectiveness of Post-Processing. The ablation studies also quantify the improvement by post-processing. Comparison in Table 5 between rows 2 & 3, 5 &6 shows that the elimination of false positives on non-leak frames significantly improves the segmentation accuracy. Fig. 7 visualizes the effectiveness of our proposed model in eliminating false positives. Previous methods (Columns 2–4) frequently produce false positives on non-leak frames. This often occurs when clouds are mistakenly identified as leaks, as seen in videos SimGas 20 & 22, due to their visual similarity. Even in non-cloud scenes, such as video SimGas 15, these

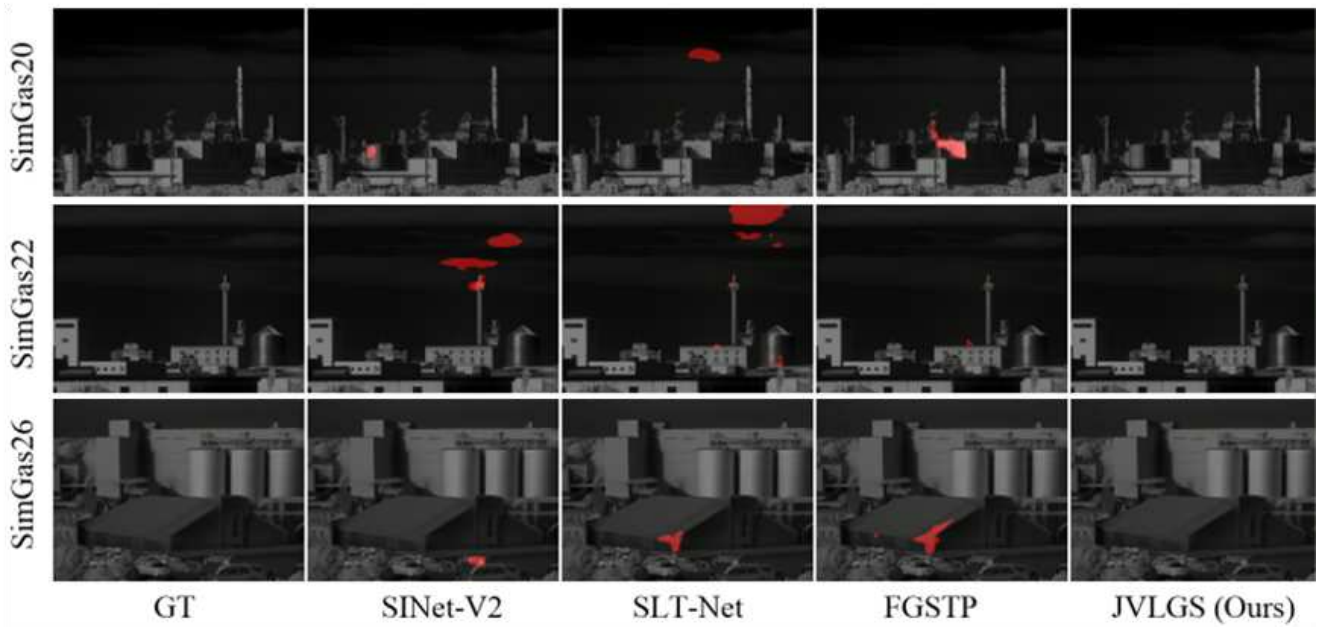


Fig. 7: False-positive removal results. Our model suppresses false positives on non-leak frames and produces masks only when leaks are present.

methods incorrectly highlight the fluctuating luminance as objects. However, the results of our model do not have false positives on non-leak frames, and the stability of model predictions has been greatly improved.

Coordination Between Modules. An interesting observation from Table 5 is that Rows 1 and 2 appear to suggest a negative impact of the VLF module on the SimGas dataset. Even when the decoder is switched to FPN in Row 5, the performance remains lower than that of Row 1, which uses the NCD decoder without VLF. This discrepancy can be attributed to interactions among components: the VLF module performs better when paired with the FPN decoder, as evidenced by the improvements between Rows 2 and 5, Rows 3 and 6. Notably, the combination of VLF, FPN, and post-processing in Row 6 achieves the highest scores on both datasets, underscoring the importance of effective coordination between modules.

4.6 Performance-Efficiency Balance

Finally, we analyze the computational efficiency of the proposed method. As shown in Fig. 8, JVLGS achieves a favorable trade-off between segmentation accuracy and processing speed. Among all benchmark models, our method attains superior performance, achieving approximately 70 $\mathcal{J}\&\mathcal{F}$, while sustaining a real-time inference speed of 40 FPS. This balanced combination of high accuracy and efficiency makes JVLGS well-suited for practical industrial monitoring applications.

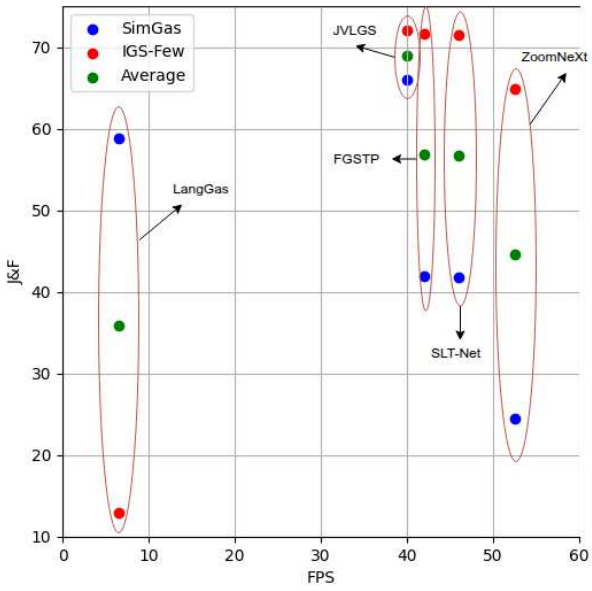


Fig. 8: Performance-Efficiency Diagram. Blue and red points represent results on our two datasets, SimGas and IGS-Few, while green points are the average accuracy across both datasets.

5 Conclusion

In this paper, we presented JVLGS, a novel framework designed to achieve high-accuracy gas leak detection and segmentation. By integrating the complementary strengths of visual features and text prompts, JVLGS enhances feature representation and significantly improves segmentation precision. Extensive experimental results demonstrate that JVLGS achieves superior mask quality under both supervised and few-shot learning settings, outperforming existing state-of-the-art methods. Furthermore, an adaptive post-processing module effectively suppresses false positives in non-leak scenarios, ensuring greater robustness and reliability in real-world industrial environments.

Overall, JVLGS offers a practical and efficient solution for industrial gas leak monitoring and safety assurance. In future work, we plan to extend JVLGS to additional video datasets and further explore its potential for real-time industrial surveillance in real world deployments.

Author Contributions Xinlong Zhao contributed to methodology development, model implementation, data curation, experiments, result analysis, and manuscript drafting. Qixiang Pang contributed to technical guidance and critical revision of the manuscript. Shan Du contributed to study conception, research guidance, result interpretation, and critical revision of the manuscript. All authors reviewed the manuscript, approved the final version, and agreed to be accountable for all aspects of the work.

Data Availability All data generated or analyzed during this study are included in this article.

Acknowledgements This work is supported by the Social Sciences and Humanities Research Council of Canada (SSHRC) under grant NFRF-GR024801.

Declaration

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Badawi, D., Pan, H., Cetin, S.C., Enis Çetin, A.: Computationally efficient spatio-temporal dynamic texture recognition for volatile organic compound (voc) leakage detection in industrial plants. IEEE Journal of Selected Topics in Signal Processing **14**(4), 676–687 (2020). DOI 10.1109/JSTSP.2020.2976555
2. Bekuzarov, M., Bermudez, A., Lee, J.Y., Li, H.: Xmem++: Production-level video segmentation from few annotated frames. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 635–644 (2023)

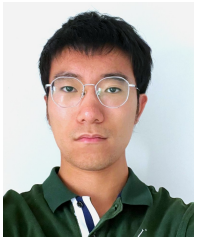
3. Cheng, H.K., Oh, S.W., Price, B., Lee, J.Y., Schwing, A.: Putting the object back into video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3151–3161 (2024)
4. Cheng, X., Xiong, H., Fan, D.P., Zhong, Y., Harandi, M., Drummond, T., Ge, Z.: Implicit motion handling for video camouflaged object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 13864–13873 (2022)
5. Fan, D.P., Ji, G.P., Cheng, M.M., Shao, L.: Concealed object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(10), 6024–6042 (2021)
6. Galun, Sharon, Basri, Brandt: Texture segmentation by multiscale aggregation of filter responses and shape elements. In: Proceedings Ninth IEEE international conference on computer vision, pp. 716–723. IEEE (2003)
7. Guo, W., Du, Y., Du, S.: Langgas: Introducing language in selective zero-shot background subtraction for semi-transparent gas leak detection with a new dataset. arXiv preprint arXiv:2503.02910 (2025)
8. He, C., Li, K., Zhang, Y., Zhang, Y., Guo, Z., Li, X., Danelljan, M., Yu, F.: Strategic preys make acute predators: Enhancing camouflaged object detectors by generating camouflaged objects. arXiv preprint arXiv:2308.03166 (2023)
9. He, S., Ding, H.: Decoupling static and hierarchical motion perception for referring video segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13332–13341 (2024)
10. Hui, W., Zhu, Z., Zheng, S., Zhao, Y.: Endow sam with keen eyes: Temporal-spatial prompt learning for video camouflaged object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19058–19067 (2024)
11. Kamath, A., Singh, M., LeCun, Y., Synnaeve, G., Misra, I., Carion, N.: Mdetr - modulated detection for end-to-end multi-modal understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 1780–1790 (2021)
12. Kopbayev, A., Khan, F., Yang, M., Halim, S.Z.: Gas leakage detection using spatial and temporal neural network model. Process Safety and Environmental Protection **160**, 968–975 (2022)
13. Lamdouar, H., Yang, C., Xie, W., Zisserman, A.: Betrayed by motion: Camouflaged object discovery via motion segmentation. In: Proceedings of the Asian conference on computer vision (2020)
14. Li, L.H., Yatskar, M., Yin, D., Hsieh, C.J., Chang, K.W.: Visualbert: A simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557 (2019)
15. Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., et al.: Grounded language-image pre-training. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 10965–10975 (2022)
16. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection (2017). URL `https://arxiv.org/abs/1612.03144`
17. Lu, Q., Li, Q., Hu, L., Huang, L.: An effective low-contrast SF_6 gas leakage detection method for infrared imaging. IEEE Transactions on Instrumentation and Measurement **70**, 1–9 (2021)
18. Minderer, M., Gritsenko, A., Houlsby, N.: Scaling open-vocabulary object detection. Advances in Neural Information Processing Systems **36**, 72983–73007 (2023)
19. Minderer, M., Gritsenko, A., Stone, A., Neumann, M., Weissenborn, D., Dosovitskiy, A., Mahendran, A., Arnab, A., Dehghani, M., Shen, Z., et al.: Simple open-vocabulary object detection. In: European conference on computer vision, pp. 728–755. Springer (2022)
20. Ouyang, L., Liu, R., Huang, Y., Furuta, R., Sato, Y.: Actionvos: Actions as prompts for video object segmentation pp. 216–235 (2024)
21. Pang, Y., Zhao, X., Xiang, T.Z., Zhang, L., Lu, H.: Zoomnext: A unified collaborative pyramid network for camouflaged object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024). DOI 10.1109/TPAMI.2024.3417329
22. Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 724–732 (2016)
23. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning, pp. 8748–8763. PmLR (2021)
24. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: International conference on machine learning, pp. 8821–8831. Pmlr (2021)
25. Ravikumar, A.P., Wang, J., Brandt, A.R.: Are optical gas imaging technologies effective for methane leak detection? Environmental science & technology **51**(1), 718–724 (2017)
26. Soille, P., Soille, P.: Opening and closing. Morphological image analysis: Principles and applications pp. 105–137 (2004)
27. Wang, J., Ji, J., Ravikumar, A.P., Savarese, S., Brandt, A.R.: Videogasnet: Deep learning for natural gas methane leak classification using an infrared camera. Energy **238**, 121516 (2022)
28. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pvt v2: Improved baselines with pyramid vision transformer. Computational visual media **8**(3), 415–424 (2022)
29. Wei, J., Wang, S., Huang, Q.: F^3net: fusion, feedback and focus for salient object detection. In: Proceedings of the AAAI conference on artificial intelligence, pp. 12321–12328 (2020)
30. Xiao, F., Hu, S., Shen, Y., Fang, C., Huang, J., He, C., Tang, L., Yang, Z., Li, X.: A survey of camouflaged object detection and beyond. arXiv preprint arXiv:2408.14562 (2024)
31. Xu, Z., Wanguo, W., Xinrui, L., Bin, L., Yuan, T.: Flame and smoke detection in substation based on wavelet analysis and convolution neural network. In: Proceedings of the 2019 3rd International Conference on Innovation in Artificial Intelligence, pp. 248–252 (2019)
32. Yan, J., Le, T.N., Nguyen, K.D., Tran, M.T., Do, T.T., Nguyen, T.V.: Mirrornet: Bio-inspired camouflaged object segmentation. IEEE Access **9**, 43290–43300 (2021)
33. Yan, P., Li, G., Xie, Y., Li, Z., Wang, C., Chen, T., Lin, L.: Semi-supervised video salient object detection using pseudo-labels. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 7284–7293 (2019)
34. Yang, C., Lamdouar, H., Lu, E., Zisserman, A., Xie, W.: Self-supervised video object segmentation by motion grouping. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 7177–7188 (2021)
35. Ye, W., Zhao, J., Wang, S., Wang, Y., Zhang, D., Yuan, Z.: Dynamic texture based smoke detection using surfacelet transform and hmt model. Fire Safety Journal **73**, 91–101 (2015)
36. Ying, S., Kunming, S., Jing, W., Feng, H., Yuze, G.: Optical gas detection: key technologies and applications review. Opto-Electronic Engineering **47**(4), 190280–1 (2020)
37. Yu, H., Wang, J., Wang, Z., Yang, J., Huang, K., Lu, G., Deng, F., Zhou, Y.: A lightweight network based on local-global feature fusion for real-time industrial invisible gas detection with infrared thermography. Applied Soft Computing **152**, 111138 (2024)
38. Yuan, J., Mao, W., Hu, C., Zheng, J., Zheng, D., Yang, Y.: Leak detection and localization techniques in oil and gas pipeline: A bibliometric and systematic review. Engineering Failure Analysis **146**, 107060 (2023)
39. Zhang, J., Huang, J., Jin, S., Lu, S.: Vision-language models for vision tasks: A survey. IEEE Trans. Pattern Anal. Mach. Intell.

46(8), 5625–5644 (2024). DOI 10.1109/TPAMI.2024.3369699. URL https://doi.org/10.1109/TPAMI.2024.3369699

40. Zhang, X., Xiao, T., Ji, G.P., Wu, X., Fu, K., Zhao, Q.: Explicit motion handling and interactive prompting for video camouflaged object detection. Trans. Img. Proc. **34**, 2853–2866 (2025). DOI 10.1109/TIP.2025.3565879. URL https://doi.org/10.1109/TIP.2025.3565879

41. Zhao, X., Du, S.: Fine-grained spatial-temporal perception for gas leak segmentation. In: 2025 IEEE International Conference on Image Processing (ICIP), pp. 869–874 (2025). DOI 10.1109/ICIP55913.2025.11084304

42. Zimmerle, D., Vaughn, T., Bell, C., Bennett, K., Deshmukh, P., Thoma, E.: Detection limits of optical gas imaging for natural gas leak detection in realistic controlled conditions. Environmental Science & Technology **54**(18), 11506–11514 (2020). DOI 10.1021/acs.est.0c01285. URL https://doi.org/10.1021/acs.est.0c01285. PMID: 32786569



Xinlong Zhao received his Master's degree in Computer Science from the University of British Columbia in 2025. He is an IEEE Student Member. He is currently a Computer Vision Algorithm Engineer at Qihoo 360 in Beijing, China. His work focuses on developing advanced computer vision algorithms and deploying large-scale AI systems for real-world applications.
His research interests include Vision–Language Models, video inpainting, and camouflaged video object detection. He has also worked on efficient AI inference architecture deployment. His broader interest lies in building practical AI systems that bridge academic research and cutting-edge industrial applications.



Qixiang Pang received his Ph.D. from Beijing University of Posts and Telecommunications and completed postdoctoral research in Electrical and Computer Engineering at the University of British Columbia. He is currently an Assistant Professor in the Department of Computer Science and Cybersecurity at the University of Central Missouri, USA. Previously, he worked at UBC, Lakehead University, and the Chinese University of Hong Kong, and in industry with Motorola, General Dynamics, and TELUS. His research focuses on computer vision, and cloud computing. He has served on technical program committees for over 50 international conferences and is an invited Editor for the Computers journal.



Shan Du is an assistant professor of Computer Science at the University of British Columbia, Okanagan. She received the Ph.D. degree in Electrical and Computer Engineering from UBC Vancouver in 2009. Before joining UBC, she worked as an assistant professor with Lakehead University and as a Research Software Engineer with IntelliView Technologies Inc. Shan has 15+ years research experience on computer vision, image and video processing, and machine learning. Shan serves as an Associate Editor of IEEE Trans. on Multimedia and IEEE Trans. on Circuits and Systems for Video Technology, Session Chair of ICIP 2019 – 2026, and TPC co-chair of GI 2025.