TRANSLATING HYPERSPECTRAL IMAGES INTO MULTIMODAL SEMANTICS THAT VISION-LANGUAGE MODELS CAN BETTER UNDERSTAND

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite capturing unified spatial-spectral representations, hyperspectral foundation models incur prohibitive pre-training costs and require sensor/scene-specific fine-tuning, severely hindering real-world deployment. This raises a fundamental question: "Can we eliminate domain-specific pre-training and adaptation, directly leveraging pre-trained vision-language models (VLMs) for training-free hyperspectral interpretation?" To address this, we propose HyperMAG, a novel framework that rethinks hyperspectral image interpretation through multimodal semantic augmented generation. Our key insight is that pre-trained VLMs, despite having never seen hyperspectral data, can reliably interpret spectral information when effectively translated into their native modalities. Specifically, HyperMAG translates spectral signatures into complementary textual descriptors (band-indexed values and peak–valley trends) and visual strips (encoding magnitude and local contrast) across multi-scale homogeneous regions, enabling off-the-shelf VLMs to extract rich spatial-spectral semantics. Extensive experiments demonstrate that HyperMAG matches or surpasses both task-specific models and foundation models across diverse downstream tasks, including few-shot classification, single-class classification, target detection, anomaly detection, and change detection.

1 INTRODUCTION

By recording continuous spectral signatures across hundreds of narrow bands, hyperspectral imagery (HSI) captures fine-grained material fingerprints inaccessible to standard RGB sensors (Veganzones et al., 2014). This detailed spectral signature explicitly encodes intrinsic physical and chemical characteristics, establishing HSI as a core modality in precision agriculture, environmental monitoring, and target detection (Bioucas-Dias et al., 2013; He et al., 2017).

Driven by the need to overcome the limitations of conventional, task-specific designs, foundation models (Achiam et al., 2023; Chang et al., 2024) have transformed the field of HSI interpretation. By leveraging masked image modeling on massive unannotated HSI datasets (Li et al., 2022; Braham et al., 2025), these models learn unified spatial-spectral representations for downstream applications (Guerri et al., 2024). However, this paradigm suffers from critical limitations. On the one hand, pre-training a foundation model from scratch incurs prohibitive computational and data collection costs. On the other hand, because current large-scale HSI datasets lack pixel-level physical semantic annotations, establishing robust cross-modal alignment remains challenging. Consequently, transferring these models to new sensors or scenes still relies on expensive, domain-specific fine-tuning (Li et al., 2025b; Liu et al., 2026).

Vision-language models (VLMs), such as CLIP (Radford et al., 2021), have shown impressive zero-shot performance in natural domain applications (Li et al., 2024a). Additionally, RemoteCLIP (Liu et al., 2024) and GeoRSCLIP (Zhang et al., 2024) have pioneered pre-training foundations for optical remote sensing imagery. However, directly applying these VLMs to HSI processing remains challenging due to a significant domain discrepancy. Unlike RGB or optical remote sensing images that contain only three channels, HSI spans hundreds of bands encoding complex physical signatures. Furthermore, HSI lacks semantic descriptions compatible with VLM text prompts. Consequently, current paradigms still depend on heavy domain pre-training or fine-tuning for HSI adaptation.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

To address these fundamental gaps, we rethink HSI adaptation from a data-centric perspective, **rather than adapting the VLM to target domain data through expensive training, we translate hyperspectral signatures into the native modalities of the frozen VLM**. Guided by this idea, we propose **HyperMAG**, a novel framework that rethinks HSI interpretation through **M**ultimodal semantic **A**ugmented **G**eneration. HyperMAG operates entirely without gradient updates by translating complex spectral signatures into VLM-native modalities. This process unfolds in three key stages. Firstly, we construct multi-scale spatially homogeneous regions to capture structural dependencies across scenes. Secondly, to bridge the spectral-semantic gap, we design a dual-branch translator that translates 1D spectral curves into structured textual descriptors capturing band-indexed values and peak-valley trends, alongside color-rendered visual strips that preserve radiometric magnitude and local contrast via global-local dual normalization. Finally, a frozen VLM extracts multimodal features from these translated representations, which are dynamically fused and aggregated across scales to significantly enhance interpretation performance. By explicitly bridging HSI physics with VLM semantics, HyperMAG eliminates domain-specific pre-training or fine-tuning.

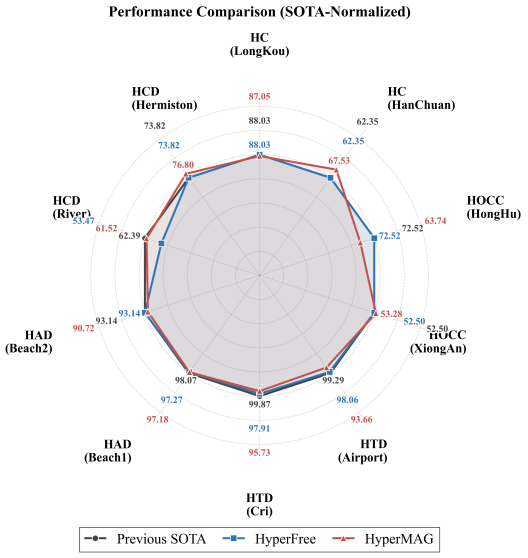

Figure 1: Without domain-specific pre-training or fine-tuning, HyperMAG achieves performance comparable to or even superior to state-of-the-art task-specific approaches and pre-trained foundation models across diverse HSI interpretation tasks.

To demonstrate the capability of HyperMAG, we comprehensively evaluate it across five hyperspectral interpretation tasks on 16 datasets. Our core innovation lies in a novel modality-translation strategy that converts 1D non-natural spectral curves into text sequences and pseudo-color images. **Remarkably, this multimodal semantics enables a completely frozen vision-language model (VLM) to achieve SOTA-level performance without any domain-specific pre-training, as shown in Fig. 1.** Crucially, the two modalities complement each other: the text representation provides the extreme numerical sensitivity required for precise target and anomaly detection, while the image representation captures global spectral trends to ensure robust classification. Ultimately, these findings confirm that unified VLMs with moderate depth can efficiently project translated multimodal representations into highly discriminative spaces, establishing a highly scalable, training-free paradigm for HSI interpretation.

2 Related Work

Task-Specific HSI Interpretation. Early hyperspectral image (HSI) interpretation relied on statistical machine learning, utilizing support vector machines (SVMs) with RBF kernels for classification (Melgani & Bruzzone, 2004a), the Reed–Xiaoli (RX) detector and its localized variants for anomaly detection (Reed & Yu, 1990; Molero et al., 2013), constrained energy minimization (CEM) for target detection (Chang, 2024), and change vector analysis (CVA) (Chen et al., 2003) alongside multivariate alteration detection (MAD) (Nielsen, 2007) for change detection. Although mathematically transparent, these methods depend on rigid statistical assumptions that limit their applicability to complex real-world scenes. This motivated the adoption of deep learning, including spatial-spectral networks for classification (Zhong et al., 2017; Roy et al., 2019), autoencoder and memory based detectors for anomaly detection (Zhang et al., 2026; Huyan et al., 2022), and siamese architectures for change detection (Ding et al., 2023; Wei et al., 2026), built upon CNNs (Krizhevsky et al., 2012), RNNs (Medsker et al., 2001), and transformers (Vaswani et al., 2017). Despite their superior performance, such approaches remain highly specialized, they generalize poorly across sensors and tasks, requiring extensive task-specific training on target-domain HSI data.

HSI Foundation Models and Pre-training Bottlenecks. Driven by the rapid emergence of large-scale datasets (Li et al., 2022; Braham et al., 2025; Zeng et al., 2026), self-supervised masked image modeling (MIM), encompassing continuous signal reconstruction (MAE (He et al., 2022)) and discrete token prediction (BEiT (Bao et al., 2021)), has become the dominant paradigm for spatial-spectral representation learning. Extending these concepts to 3D HSI cubes, early works designed tailored backbones such as FactoFormer (Mohamed et al., 2023), HSIMAE (Wang et al., 2024), S2HM2 (Tu et al., 2024), SpectralMAE (Zhu et al., 2023), and S2MAE (Li et al., 2024b) to capture invariant spatial structures and localized spectral continuity for downstream fine-tuning. Building upon these pre-training designs, recent efforts have scaled MIM paradigms into universal HSI foundation models, such as SpectralGPT (Hong et al., 2024) and HyperSIGMA (Wang et al., 2025), enabling unified feature extraction across diverse sensors and scenes. Despite these advances, deploying current hyperspectral foundation models faces two critical bottlenecks: first, task-specific annotated HSI data remains remarkably scarce, severely restricting effective fine-tuning; and second, pre-training such foundation architectures from scratch incurs prohibitive computational costs and massive data collection overheads.

Vision-Language Models and Training-Free Adaptation. Vision-language models (VLMs) such as CLIP (Radford et al., 2021), SigLIP (Zhai et al., 2023), and Qwen-VL (Li et al., 2026) align visual and textual semantics in a unified latent space. To deploy these foundation models without fine-tuning overhead, training-free adaptation has progressively evolved from textual enrichment to visual-level manipulation. At the textual level, LLM-generated descriptors replace generic class names with fine-grained physical attributes (Pratt et al., 2023; Menon & Vondrick, 2022). At the feature level, non-parametric support caches refine prediction distributions via metric tuning without backpropagation (Zhang et al., 2021; 2023). At the visual input level, task-specific spatial cues are directly injected into image streams via visual prompting (Stein et al., 2025). By eliminating gradient updates, these parameter-free methods preserve foundation model priors while enabling highly efficient cross-domain transfer. However, directly extending training-free VLMs to hyperspectral imagery encounters a fundamental modality gap: standard VLMs are trained exclusively on 3-channel RGB data, and generic prompts cannot express physical spectral signatures. Addressing this challenge, HyperMAG rethinks HSI interpretation through multimodal semantic translation, grounded in the insight that frozen VLMs can reliably interpret spectral physics when converted into structured textual descriptors and color-rendered visual strips. This data-level adaptation enables off-the-shelf VLMs to extract rich spatial-spectral semantics for parameter-free, zero-shot HSI interpretation without any domain-specific pre-training.

3 METHOD

Overview. Given a hyperspectral image $\mathbf{X} \in \mathbb{R}^{B\times H\times W}$, HyperMAG eliminates the need for gradient updates by translating spectral curves into the native modalities of a frozen VLM. As shown in Fig. 2, HyperMAG contains three stages: multi-scale homogeneous region generation, multimodal spectral translation, and feature extraction with fusion. Formally, for an aggregated regional spectrum $\mathbf{x} \in \mathbb{R}^B$, dual-branch translators ϕ_{txt} and ϕ_{img} map the signal into the text space $\mathcal{T}$ and image space $\mathcal{I}$.

$$\phi_{\text{txt}} : \mathbb{R}^B \to \mathcal{T}, \quad \phi_{\text{img}} : \mathbb{R}^B \to \mathcal{I}. \tag{1}$$

Then, we employ the frozen VLM to extract complementary semantic features from these translated spaces, yielding training-free spatial-spectral representations for downstream interpretation tasks.

$$f_{\text{txt}} = \mathcal{F}_{\text{txt}}\big(\phi_{\text{txt}}(\mathbf{x})\big), \quad f_{\text{img}} = \mathcal{F}_{\text{img}}\big(\phi_{\text{img}}(\mathbf{x})\big), \tag{2}$$

where $\mathcal{F}_{\text{txt}}$ and $\mathcal{F}_{\text{img}}$ are the text encoder and visual encoder of the pre-trained VLM, respectively.

Multi-Scale Homogeneous Region Generation. To exploit spatial context and mitigate pixel-level noise, HyperMAG aggregates spectral signatures within structurally homogeneous regions. We first apply Principal Component Analysis (PCA) (Abdi & Williams, 2010) on $\mathbf{X}$ to extract a 3-channel pseudo-color image. Next, we generate region label maps $\mathbf{L}^{(c)} \in \mathbb{N}^{H\times W}$ using Felzenszwalb segmentation (Felzenszwalb & Huttenlocher, 2004) across multiple scale parameters $c \in \{c_1, \ldots, c_C\}$:

$$\mathbf{L}^{(c)} = \text{Felz}\big(\text{PCA}(\mathbf{X}); c\big). \tag{3}$$

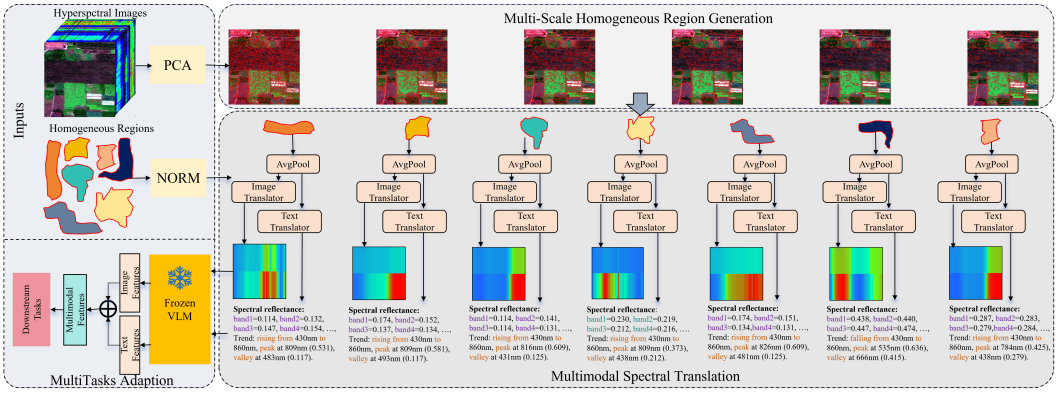

Figure 2: **Overall architecture of HyperMAG.** The training-free framework operates in three key stages: (1) **Multi-Scale Region Generation:** HSI cubes undergo PCA dimension reduction and Felzenszwalb superpixel segmentation to extract spatially homogeneous regions across multiple scales. (2) **Multimodal Spectral Translation:** Regional spectral pixels are encoded into complementary text descriptors (band-indexed values with peak-valley trends) and visual strips (global-local normalized color strips) to bridge the spectral-semantics gap. (3) **Multi-Task Adaptation:** A frozen VLM extracts and fuses multimodal features into unified regional representations, which are ensembled across scales to be applied to five downstream interpretation tasks without any domain pre-training or fine-tuning.

For each region $r \in \{1, \ldots, R\}$ in $\mathbf{L}^{(c)}$, we calculate its average pixel along the spatial dimension as the representative spectrum $\bar{\mathbf{x}}_r$:

$$\bar{\mathbf{x}}_r = \frac{1}{|\mathcal{R}_r|} \sum_{(h,w)\in\mathcal{R}_r} \mathbf{x}_{hw}, \tag{4}$$

where $\mathcal{R}_r$ denotes the set of homogeneous spectral pixels in region r. Multi-scale region segmentation avoids single-scale bias, yielding robust inputs for subsequent modality translation. To simplify the formula expression, we omitted the region indicators in the subsequent multimodal spectral translation and multimodal feature extraction and fusion sections.

Multimodal Spectral Translation. Before spectral translation, to unify the variable radiometric ranges across sensors, we apply Min-Max normalization to the region spectrum and achieve the normalized spectrum, $\tilde{\mathbf{x}}$.

Spectral-Text Translation. Standard VLMs cannot directly perceive physical spectral curves, but show strong capability in processing numerical sequences (Li et al., 2025a; Wu et al., 2026). Leveraging this prior, we convert the normalized spectrum $\tilde{\mathbf{x}}$ and its wavelength sequence $\mathbf{w} \in \mathbb{R}^B$ into a structured textual descriptor $\phi_{\text{txt}}(\tilde{\mathbf{x}}, \mathbf{w})$. As shown the examples in Fig. 3, the text descriptor concatenates band-indexed values with a global trend analysis:

$$\phi_{\text{txt}}(\tilde{\mathbf{x}}, \mathbf{w}) = \text{“Spectral reflectance: ”} \oplus \text{Seq}(\tilde{\mathbf{x}}) \oplus \text{“ Trend: ”} \oplus \tau(\tilde{\mathbf{x}}, \mathbf{w}), \tag{5}$$

where $\text{Seq}(\tilde{\mathbf{x}})$ formats each band value into a 3-decimal string, and $\oplus$ denotes string concatenation. The trend function $\tau(\tilde{\mathbf{x}}, \mathbf{w})$ computes inter-band gradients to determine the overall slope (“rising”, “falling”, or “stable”) and explicitly logs global reflectance peaks and valleys alongside their respective wavelengths. This enhances the numerical sensitivity of $\mathcal{F}_{\text{txt}}$, facilitating fine-grained material discrimination.

Spectral-Image Translation. To satisfy the input requirements of the VLM visual encoder $\mathcal{F}_{\text{img}}$, the visual translator $\phi_{\text{img}}(\tilde{\mathbf{x}})$ linearly interpolates the 1D spectrum into a 224×224 visual strip. To simultaneously capture absolute radiometric magnitude and subtle spectral variations, we employ a global-local dual normalization strategy, as illustrated in Fig. 3. The normalized spectrum is mapped into RGB space and split vertically into two regions. The upper half encodes global normalization scaled relative to the entire dataset extrema to facilitate magnitude comparison across different spectra. Meanwhile, the lower half encodes local normalization scaled relative to the individual spectrum extrema to enhance local shape contrast.

Multimodal Feature Extraction and Fusion. Finally, the frozen VLM extracts rich semantic embeddings from the translated textual descriptors and visual strips. The normalized multimodal se-

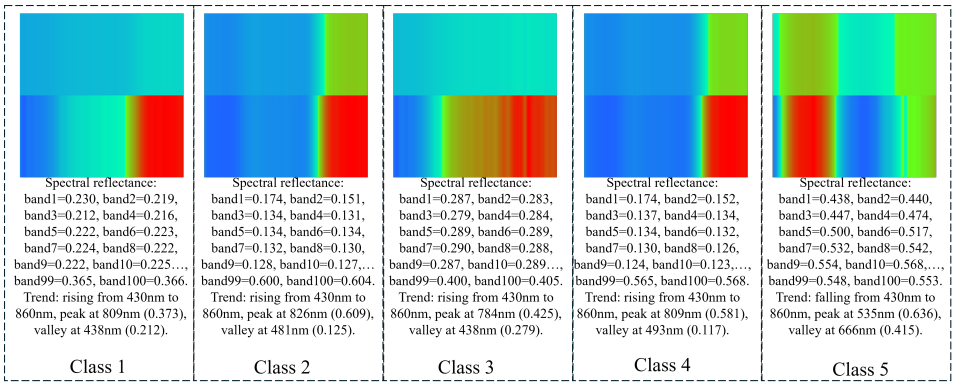

Figure 3: Textual and visual representations obtained via the text-translator and image-translator for five categories on the PaviaU dataset.

mantical features can be represented as follows:

$$\mathbf{z}_{\text{txt}} = ||\mathcal{F}_{\text{txt}}\big(\phi_{\text{txt}}(\tilde{\mathbf{x}}, \mathbf{w})\big)||_2, \quad \mathbf{z}_{\text{img}} = ||\mathcal{F}_{\text{img}}\big(\phi_{\text{img}}(\tilde{\mathbf{x}})\big)||_2. \tag{6}$$

Since text descriptors provide precise numerical bounds while visual strips capture continuous spectral shapes, we fuse the multimodal features to capture complementary information through the channel fusion operator:

$$\mathbf{z} = \big[\mathbf{z}_{\text{txt}} \parallel \mathbf{z}_{\text{img}}\big] \in \mathbb{R}^{2d}, \tag{7}$$

where d is the latent feature dimension. These unified representations are directly deployed for zero-shot downstream tasks without requiring further domain adaptation.

Downstream Task Applications. The domain-agnostic multimodal representations extracted by HyperMAG enable direct, zero-shot deployment across a wide range of hyperspectral interpretation tasks in a training-free manner. Specifically, we evaluate HyperMAG on five representative downstream applications: few-shot classification (HC) (Liu et al., 2018), one-class classification (HOCC) (Zhao et al., 2023b), target detection (HTD) (Chen et al., 2023), change detection (HCD) (Liu et al., 2019), and anomaly detection (HAD) (Xu et al., 2022).

For classification, target detection, and change detection, the inference process is formulated as metric matching, where query region features are evaluated against target prototypes or reference centroids via cosine similarity or hybrid semantic-physical distance. For anomaly detection, we extend the classic RX operator (Reed & Yu, 1990) into the VLM feature space using regularized Mahalanobis distance, fusing semantic distinctiveness with raw spectral variations. Pixel-level predictions across all tasks are finalized by aggregating region-level results across multiple scales to mitigate single-scale segmentation bias. Detailed mathematical formulations and implementation procedures for each task are provided in Appendix A.

4 Experimental Results

4.1 Experimental Setting

Datasets. To comprehensively evaluate our method, we employ diverse benchmark datasets specific to each task. **HC**: Indian Pines (Baumgardner et al., 2015), Pavia University (Huang & Zhang, 2009), and WHU-Hi (LongKou, HanChuan, HongHu) (Zhong et al., 2020). **HOCC**: WHU-Hi-HongHu (Zhong et al., 2020) and XiongAn (Cen et al., 2020). **HTD**: ABU-airport-4 (Kang, 2017) and the Cri dataset. **HAD**: 13 sub-scenes from the ABU dataset (Kang, 2017) and the HYDICE dataset (Li & Du, 2014). **HCD**: River and Hermiston bitemporal datasets (Hu et al., 2024). Detailed descriptions and preprocessing protocols for these datasets are provided in Appendix E.1.

Implementation Details. We employ the Qwen3-VL-Embedding-2B (Li et al., 2026) model as the foundation of HyperMAG. During all experiments, the model parameters remain completely frozen without any fine-tuning. For multi-scale spatial modeling, we apply the Felzenszwalb superpixel

segmentation algorithm using six scale parameters set to $\{0.08, 0.18, 0.30, 0.45, 0.55, 0.70\} \times \alpha_s$, where $\alpha_s = \min(H, W)$. For the **text branch**, we design a structured prompt that explicitly integrates sequential band values with physical spectral statistics. In the **image branch**, a pseudo-color mapped spectral strip image is obtained by global normalization and single-pixel spectral normalization of spectral strips, and its size is 224×224.

4.2 Multi-Task Evaluation and Comparative Analysis

Table 1: Quantitative comparison on five HSI interpretation tasks. The best and second-best results on each dataset are highlighted with dark and light red backgrounds, respectively.

HC (Few-shot Classification)

Metric	LongKou						HanChuan					
	SVM (5-shot)	HybridSN (5-shot)	SpecFormer (5-shot)	SSFTT (5-shot)	HyperFree (1-pmt)	HyperMAG (5-shot)	RBF-SVM (5-shot)	HybridSN (5-shot)	SpecFormer (5-shot)	SSFTT (5-shot)	HyperFree (1-pmt)	HyperMAG (5-shot)
OA	79.07	82.23	82.23	88.43	93.39	89.03 (2nd)	54.51	42.67	43.26	59.98	75.47	71.80 (2nd)
AA	70.35	51.91	51.91	78.88	88.03	87.05 (2nd)	46.36	26.48	26.46	46.23	62.35	67.53 (1st)
KA	0.74	0.77	0.77	0.85	0.91	0.86 (2nd)	0.48	0.28	0.32	0.53	0.72	0.67 (2nd)

HOCC (One-class Classification)

Metric	HongHu						XiongAn					
	OCSVM (5-shot)	BSVM (5-shot)	PAN (5-shot)	T-HOC (5-shot)	HyperFree (1-pmt)	HyperMAG (5-shot)	OCSVM (5-shot)	BSVM (5-shot)	PAN (5-shot)	T-HOC (5-shot)	HyperFree (1-pmt)	HyperMAG (5-shot)
F_1	26.33	34.82	63.69	55.97	72.52	63.74 (2nd)	18.31	26.30	46.34	41.34	52.50	53.28 (1st)
P	56.43	50.77	80.00	46.52	68.61	71.60 (2nd)	39.83	23.82	47.13	32.87	68.21	58.07 (2nd)
R	24.02	45.29	64.27	92.35	78.06	71.25 (3rd)	16.08	57.83	53.32	60.38	65.03	73.17 (1st)

HTD (Supervised Target Detection)

Metric	Airport						Cri					
	ACE (1-shot)	CEM (1-shot)	MF (1-shot)	HTD-IRN (1-shot)	HyperFree (1-pmt)	HyperMAG (1-shot)	ACE (1-shot)	CEM (1-shot)	MF (1-shot)	HTD-IRN (1-shot)	HyperFree (1-pmt)	HyperMAG (1-shot)
AUROC	97.94	96.03	99.16	97.45	98.06	93.66 (6th)	97.35	98.93	98.91	99.75	97.91	95.73 (6th)

HAD (Unsupervised Anomaly Detection)

Metric	Beach1						Beach2					
	RX	CRD	Auto-AD	HyperSIGMA	HyperFree (0-shot)	HyperMAG	RX	CRD	Auto-AD	HyperSIGMA	HyperFree (0-shot)	HyperMAG
AUROC	98.07	97.37	96.42	96.02	97.27	97.18 (4th)	91.06	83.48	93.03	90.64	93.14	90.72 (4th)

HCD (Change Detection)

Metric	River						Hermiston					
	FC-Sc (5-shot)	FC-Sd (5-shot)	BIT (5-shot)	SST-F (5-shot)	HyperFree (0-shot)	HyperMAG (5-shot)	FC-Sc (5-shot)	FC-Sd (5-shot)	BIT (5-shot)	SST-F (5-shot)	HyperFree (0-shot)	HyperMAG (5-shot)
IoU	45.22	45.34	21.26	40.96	36.49	44.42 (3rd)	37.76	48.73	52.57	53.61	58.51	62.34 (1st)
F_1	62.28	62.39	35.07	58.12	53.47	61.52 (3rd)	54.82	65.52	68.91	69.80	73.82	76.80 (1st)

To evaluate the universal representation capability of HyperMAG, we conduct comprehensive comparisons across five distinct hyperspectral interpretation tasks. Benchmark baselines include: **HC:** SVM (Cortes & Vapnik, 1995; Melgani & Bruzzone, 2004b), HybridSN (Roy et al., 2019), SpecFormer (Hong et al., 2022), SSFTT (Sun et al., 2022), and HyperFree Li et al. (2025b). **HOCC:** OCSVM (Schölkopf et al., 1999), BSVM (Kiryo et al., 2017), PAN (Kiryo et al., 2017), T-HOC (Zhao et al., 2023a), and HyperFree (Li et al., 2025b). **HTD:** ACE (Kraut et al., 2001), CEM (Chang, 2024), MF (Manolakis et al., 2013), HTD-IRN (Shen et al., 2023), and HyperFree (Li et al., 2025b). **HAD:** RX (Reed & Yu, 1990), CRD (Li & Du, 2015), Auto-AD (Wang et al., 2022a), HyperSIGMA (Wang et al., 2025), and HyperFree Li et al. (2025b). **HCD:** FC-Sc, FC-Sd (Daudt et al., 2018), BIT (Chen et al., 2021), SST-F (Wang et al., 2022b), and HyperFree (Li et al., 2025b).

Overall Performance Across Five Tasks. As quantitative results in Table 1 and the multi-task radar profile in Fig. 1 demonstrate, HyperMAG achieves superior or highly competitive performance across all five downstream tasks without any task-specific fine-tuning or domain pre-training. In supervised and few-shot settings (HC, HCD, and HOCC), where traditional deep learning architectures severely degrade due to label scarcity, HyperMAG consistently achieves superior performance, outperforming dedicated task-specific baselines and exceeding or matching HyperFree across various benchmarks. Particularly on the HCD task, HyperMAG achieves a new state-of-the-art result (76.80% F_1 score on Hermiston). In unsupervised and one-shot scenarios (HAD and HTD), HyperMAG exhibits robust cross-domain transferability. Without requiring any downstream training or task-specific fine-tuning, HyperMAG achieves detection performance highly comparable to pre-trained HSI foundation models such as HyperFree and HyperSIGMA. Detailed dataset-level results and visual comparisons are refer to Appendix C.

Our multi-task empirical results demonstrate that mapping complex 1D continuous spectral curves into natural language descriptors and visual strips effectively bridges the spectral-semantics gap, un-

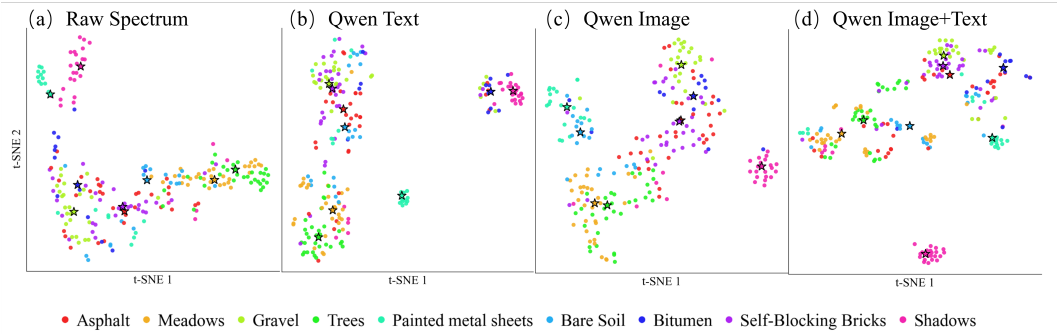

Figure 4: T-SNE visualization of feature representations on the Pavia University dataset. From left to right: raw spectrum, text features, image features, and fused image-text features (stars denote class prototypes). Compared to raw spectrum, both text and image representations significantly improve class separability by preserving fine-grained physical similarities. Their fusion further achieves high intra-class compactness and optimal discriminability.

Table 2: Ablation study of the text and image translation. **Bold** and underlined values denote the best and second-best performance, respectively.

Setting	Text	Image	HC			HOCC			HTD	HAD	HCD	
			OA	AA	KA	F_1	Precision	Recall	AUROC	AUROC	F_1	IoU
Text	×	✓	77.78	73.13	0.74	48.55	51.33	71.03	**96.61**	**88.73**	66.12	51.72
Image	✓	×	80.40	**77.56**	**0.77**	54.17	59.52	64.79	96.28	87.85	47.97	32.07
Text+Image	✓	✓	**80.42**	77.29	**0.77**	**58.51**	**64.84**	**72.21**	95.73	88.14	**69.16**	**53.38**

locking the rich prior knowledge embedded within frozen Vision-Language Models (VLMs) without requiring large-scale hyperspectral pre-training. Crucially, as shown in Fig. 4, this cross-modal translation preserves fine-grained physical similarities, yielding high intra-class compactness and iter-class separability within the shared embedding space even in the absence of explicit metric learning or fine-tuning. Consequently, HyperMAG mitigates overfitting in label-scarce scenarios and generalizes robustly across diverse sensors, wavelengths, and scenes without retraining, substantially lowering the deployment barrier for remote sensing foundation models.

5 Ablations and Analysis

Ablation on Multimodal Translators. To evaluate the contributions of the text and image branches in HyperMAG, we compare single-branch variants with the full dual-branch architecture across all five hyperspectral tasks. As observed in Table 2, jointly utilizing both branches achieves superior performance on HC, HOCC, and HCD, demonstrating strong multi-task complementarity. Interestingly, the text-only branch achieves the best results on fine-grained detection tasks (HTD and HAD). This is because textual numerical representations effectively preserve subtle spectral variations critical for isolating small targets and anomalies, whereas image representations tend to lose fine-grained details. Conversely, the image branch captures global spectral trends and spatial structure, yielding better generalization than the text branch on classification tasks. Overall, combining both modalities ensures optimal generalization across diverse hyperspectral interpretation scenarios.

Influence of Pre-trained VLMs. We evaluate HyperMAG across four representative pre-trained VLM backbones: CLIP ViT-B/16 (Radford et al., 2021), SigLIP 2 SoViT-400M (Zhai et al., 2023), Qwen3-VL-Embedding-2B (Li et al., 2026), and Qwen3-VL-Embedding-8B (Li et al., 2026). Experiments are conducted under the 5-shot classification setting on the LongKou and HanChuan datasets, with performance reported in Table 3. As observed in Table 3, Qwen3-VL-Embedding-2B and SigLIP 2 SoViT-400M achieve superior classification performance compared to CLIP ViT-B/16 and Qwen3-VL-Embedding-8B. We attribute these differences to two main reasons: 1) CLIP ViT-B/16 performs poorly due to its strict 77-token context limit, which fails to encode fine-grained

Table 3: Hyperspectral image classification results based on various VLMs. **Bold** and underlined values denote the best and second-best performance, respectively.

VLMs	LongKou			HanChuan		
	OA	AA	Kappa	OA	AA	Kappa
CLIP ViT-B/16	84.91	83.94	0.81	68.82	64.61	0.64
SigLIP 2 SoViT-400M	**90.72**	81.70	**0.88**	67.96	61.86	0.63
Qwen3-VL-2B	89.03	**87.05**	0.86	**71.80**	**67.53**	**0.67**
Qwen3-VL-8B	83.41	75.90	0.78	61.85	57.20	0.56

Spectral Values	Spectral Values + Spectral Indices	Spectral Descriptions
Class1: sp 48,91,107,128,137,140,157,194,201,189,181,183,183,173,178,203,254,427,422,519,469,344,433,463,462,311,280,108,146,285…	**Class1:** Spectral reflectance: band1=0.048, band2=0.059, band3=0.071, band4=0.079, band5=0.086, band6=0.091…Trend: rising from 402nm to 999nm, peak at 782nm (0.519), valley at 402nm (0.048).	**Class1:** Spectrum B270 median=0.2 p5=0.1 p95=0.5 dispersion=1.11 trend=rising(1.459). segments(n=6): s0=0.1 s1=0.2 s2=0.2 s3=0.4 s4=0.4 s5=0.2. seg_ratio=2.25. peaks: b177=0 b211=0 b70=0. valleys: b192=0 b248=0 b124=0.
Class3: sp 51,95,108,126,128,129,168,252,258,205,177,160,148,130,129,198,399,731,703,837,747,530,661,694,687,454,398,139,192,390…	**Class3:** Spectral reflectance: band1=0.051, band2=0.063, band3=0.076, band4=0.083, band5=0.090, band6=0.095…Trend: rising from 402nm to 999nm, peak at 755nm (0.852), valley at 402nm (0.051).	**Class3:** Spectrum B270 median=0.2 p5=0.1 p95=0.8 dispersion=1.84 trend=rising(2.172). segments(n=6): s0=0.1 s1=0.2 s2=0.1 s3=0.6 s4=0.7 s5=0.3. seg_ratio=3.47. peaks: b176=1 b210=1 b68=0. valleys: b192=1 b124=0 b248=0.
Class5: sp 76,140,169,208,223,226,252,310,326,313,303,303,303,284,293,319,373,547,503,603,538,388,483,512,504,338,301,116,158,309…	**Class5:** Spectral reflectance: band1=0.076, band2=0.094, band3=0.111, band4=0.121, band5=0.133, band6=0.140, …Trend: rising from 402nm to 999nm, peak at 755nm (0.615), valley at 402nm (0.076).	**Class5:** Spectrum B270 median=0.3 p5=0.1 p95=0.6 dispersion=0.58 trend=rising(0.677). segments(n=6): s0=0.2 s1=0.3 s2=0.3 s3=0.5 s4=0.5 s5=0.3. seg_ratio=1.60. peaks: b176=1 b210=1 b72=0. valleys: b248=0 b192=0 b125=0.

Figure 5: Visual comparison of different spectral text generation methods on the LongKou dataset.

numerical sequences across full spectral bands. In contrast, models supporting longer contexts (e.g., 256 tokens in SigLIP 2 and up to 8192 tokens in Qwen3-VL) preserve critical spectral details. 2)Since VLMs are pre-trained on natural images and text, inputting abstract numerical sequences and color-coded spectral strips introduces noticeable domain shift. The deeper text tower in Qwen3-VL-8B causes feature over-dispersion under such unnatural inputs. Conversely, the shallower architecture of Qwen3-VL-2B is less sensitive to fine-grained numerical variations, yielding stronger domain adaptability. Overall, Qwen3-VL-Embedding-2B delivers the most balanced and stable performance across datasets, and is thus selected as the default VLM backbone for HyperMAG.

Influence of Text Generation Strategies. We evaluate three text generation strategies on the text-only branch under the 5-shot classification setting: (1) *Spectral Values Only*, (2) *Spectral Values with Band Index*, and (3) *Spectral Description*. The prompt examples and classification results are presented in Fig. 5 and Table 4, respectively. Experimental results show that *Spectral Values Only* performs worst because omitting band indices discards critical positional geometry. *Spectral Description* achieves the highest accuracy by incorporating hand-crafted statistical features and shape summaries. However, complex hand-crafting hinders cross-task generalization. To balance accuracy and generality, *Spectral Values with Band Index* combines raw reflectance with band indices and key global landmarks (e.g., peaks and trends). This strategy retains maximum raw spectral fidelity and domain-agnostic simplicity while achieving highly competitive accuracy. Consequently, HyperMAG adopts *Spectral Values with Band Index* as the default text generation strategy.

Influence of Image Generation Strategies. We evaluate three image generation strategies on the image-only branch under the 5-shot classification setting: (1) *Spectral Curve Plot*, (2) *Global-Normalized Strip*, and (3) *Dual-Normalized Strip (Global + Local)*. Representative image visual examples and classification results are presented in Fig. 6 and Table 5, respectively. Experimental results indicate that while *Spectral Curve Plot* offers an intuitive 1D profile representation, it is less effective for standard vision backbones to extract dense spectral semantics. *Global-Normalized Strip* maps reflectance into a color-coded strips, capturing absolute signal magnitude for direct cross-pixel comparison. To further highlight fine-grained spectral fluctuations, *Dual-Normalized Strip* concate-

Table 4: The influence of the text generation method. **Bold** values denote the best.

Text generalization method	LongKou			HanChuan		
	OA	AA	Kappa	OA	AA	Kappa
Spectral Values	85.12	79.56	0.81	67.65	63.36	0.62
Spectral Values+ Band Index	86.90	82.62	0.83	69.05	61.13	0.64
Spectral Descriptor	**89.88**	**84.24**	**0.87**	**69.87**	62.85	**0.65**

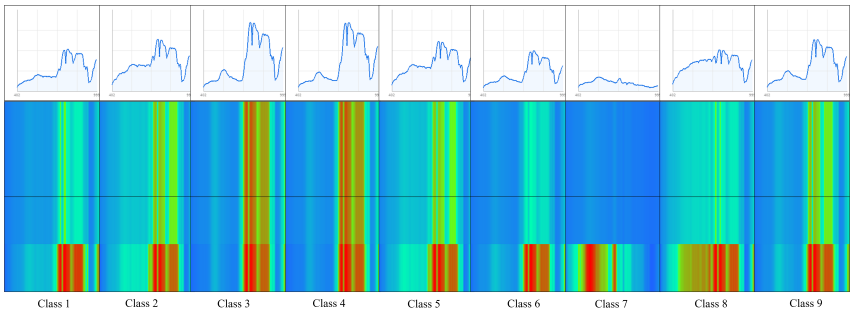

Figure 6: Visual comparison of different spectral image generation methods on the LongKou dataset.

Table 5: The influence of the image generation method.

Image generation	LongKou			HanChuan		
	OA	AA	Kappa	OA	AA	Kappa
Spectral Curve Plot	90.46	84.08	0.88	66.79	63.58	0.62
Global-Normalized Strip	83.99	84.17	0.80	69.96	65.22	0.65
Dual-Normalized Strip	**91.63**	**86.52**	**0.89**	**71.67**	**66.85**	**0.67**

nates the globally normalized strips (reflecting absolute intensity) with a locally normalized bar (amplifying subtle shape variations). This representation achieves the best accuracy across both datasets by seamlessly unifying global intensity with fine-grained spectral details. Consequently, HyperMAG adopts *Dual-Normalized Strip* as the default image generation strategy.

Influence of Multi-Scale Superpixel Segmentation. We evaluate the multi-scale superpixel ensemble strategy on the LongKou dataset under the 5-shot setting by progressively integrating up to ten scales (ranging from 32 to 280). As depicted in Fig. 7, all performance metrics (OA, AA, and Kappa) improve consistently as more scales are ensembled. This gain stems from the multi-scale aggregation, which effectively captures spatial homogeneity across diverse spatial scales to yield purer, more discriminative averaged spectrum while suppressing local noise. To balance classification accuracy with inference efficiency, HyperMAG adopts a 6-scale ensemble by default.

6 Conclusion

We introduce HyperMAG, a novel training-free framework that fundamentally rethinks hyperspectral image (HSI) interpretation. Instead of relying on computationally prohibitive domain-specific pre-training or fine-tuning, HyperMAG bridges the gap between HSI spectral signatures and foundation model semantics from a data-centric perspective. By converting complex 1D spectral curves into VLM-native modalities, such as band-indexed textual descriptors and dual-normalized visual strips, HyperMAG enables frozen Vision-Language Models (VLMs) to extract highly discriminative spatial-spectral features. Extensive experiments across five diverse interpretation tasks demonstrate that this unified paradigm matches or surpasses both specialized architectures and HSI foundation models, all without any training. Importantly, ablation studies reveal strong modality complementarity between text and visual. The text representation is highly sensitive to numerical values, which is vital for fine-grained target and anomaly detection, whereas the visual strip captures global spectral trends to ensure robust classification. This confirms that pre-trained VLMs can effectively interpret HSI data when properly translated into their native embedding spaces.

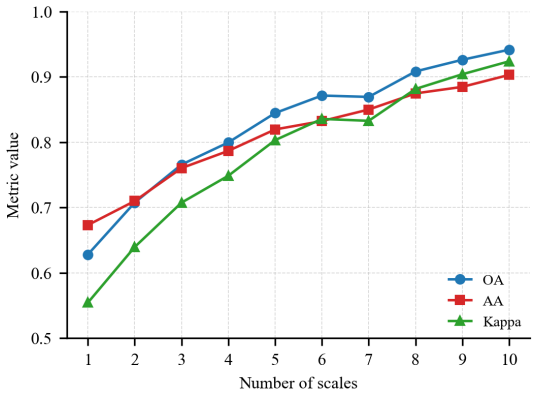

Figure 7: The classification performance by progressively integrating up to ten scales superpixel segmentation on the LongKou dataset under the 5-shot setting.

AI ASSISTANCE STATEMENT

We declare that Artificial Intelligence (AI) tools, specifically Large Language Models (LLMs), were utilized solely for text polishing, grammatical editing, and code debugging during the preparation of this manuscript. All core scientific contributions, including methodology design, mathematical formulations, experimental setup and implementation, quantitative and qualitative evaluations, and final conclusions, were developed, conducted, and authored entirely by the authors without the use of AI generation.

REFERENCES

Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

Marion Baumgardner, Larry Biehl, and David Landgrebe. 220 band aviris hyperspectral image data set: June 12, 1992 indian pine test site 3. *(No Title)*, 2015.

Yoshua Bengio and Yann LeCun. Scaling learning algorithms towards AI. In *Large Scale Kernel Machines*. MIT Press, 2007.

José M Bioucas-Dias, Antonio Plaza, Gustavo Camps-Valls, Paul Scheunders, Nasser Nasrabadi, and Jocelyn Chanussot. Hyperspectral remote sensing data analysis and future challenges. *IEEE Geoscience and remote sensing magazine*, 1(2):6–36, 2013.

Nassim Ait Ali Braham, Conrad M Albrecht, Julien Mairal, Jocelyn Chanussot, Yi Wang, and Xiao Xiang Zhu. Spectralearth: Training hyperspectral foundation models at scale. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 18:16780–16797, 2025.

Yi Cen, Lifu Zhang, Xia Zhang, Yueming Wang, Wenchao Qi, Senlin Tang, and Peng Zhang. Airborne hyperspectral remote sensing image classification dataset of matiwan village in xiong'an. *Journal of Remote Sensing*, 24(11):1299–1306, 2020.

Chein-I Chang. Constrained energy minimization (cem) for hyperspectral target detection: Theory and generalizations. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–21, 2024.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45, 2024.

Bowen Chen, Liqin Liu, Zhengxia Zou, and Zhenwei Shi. Target detection in hyperspectral remote sensing image: Current status and challenges. *Remote Sensing*, 15(13):3223, 2023.

Hao Chen, Zipeng Qi, and Zhenwei Shi. Remote sensing image change detection with transformers. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2021.

Jin Chen, Peng Gong, Chunyang He, Ruiliang Pu, and Peijun Shi. Land-use/land-cover change detection using improved change-vector analysis. *Photogrammetric Engineering & Remote Sensing*, 69(4):369–379, 2003.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

Rodrigo Caye Daudt, Bertr Le Saux, and Alexandre Boulch. Fully convolutional siamese networks for change detection. In *2018 25th IEEE international conference on image processing (ICIP)*, pp. 4063–4067. IEEE, 2018.

Jigang Ding, Xiaorun Li, Shu Xiang, and Shuhan Chen. Multilevel features fused and change information enhanced neural network for hyperspectral image change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–13, 2023.

Pedro F Felzenszwalb and Daniel P Huttenlocher. Efficient graph-based image segmentation. *International journal of computer vision*, 59(2):167–181, 2004.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT Press, 2016.

Mohamed Fadhlallah Guerri, Cosimo Distante, Paolo Spagnolo, Fares Bougourzi, and Abdelmalik Taleb-Ahmed. Deep learning techniques for hyperspectral image analysis in agriculture: A review. *ISPRS Open Journal of Photogrammetry and Remote Sensing*, 12:100062, 2024.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 15979–15988. IEEE, 2022.

Lin He, Jun Li, Chenying Liu, and Shutao Li. Recent advances on spectral–spatial hyperspectral image classification: An overview and new guidelines. *IEEE Transactions on Geoscience and Remote Sensing*, 56(3):1579–1597, 2017.

Geoffrey E. Hinton, Simon Osindero, and Yee Whye Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18:1527–1554, 2006.

Danfeng Hong, Zhu Han, Jing Yao, Lianru Gao, Bing Zhang, Antonio Plaza, and Jocelyn Chanussot. Spectralformer: Rethinking hyperspectral image classification with transformers. *IEEE Trans. Geosci. Remote Sens.*, 60:1–15, 2022. DOI: 10.1109/TGRS.2021.3130716.

Danfeng Hong, Bing Zhang, Xuyang Li, Yuxuan Li, Chenyu Li, Jing Yao, Naoto Yokoya, Hao Li, Pedram Ghamisi, Xiuping Jia, et al. Spectralgpt: Spectral remote sensing foundation model. *IEEE transactions on pattern analysis and machine intelligence*, 46(8):5227–5244, 2024.

Meiqi Hu, Chen Wu, and Liangpei Zhang. Globalmind: Global multi-head interactive self-attention network for hyperspectral change detection. *ISPRS Journal of Photogrammetry and Remote Sensing*, 211:465–483, 2024.

Xin Huang and Liangpei Zhang. A comparative study of spatial approaches for urban mapping using hyperspectral ROSIS images over Pavia city, northern italy. *International Journal of Remote Sensing*, 30(12):3205–3221, 2009. doi: 10.1080/01431160802559046.

Ning Huyan, Xiangrong Zhang, Dou Quan, Jocelyn Chanussot, and Licheng Jiao. Aud-net: A unified deep detector for multiple hyperspectral image anomaly detection via relation and few-shot learning. *IEEE Transactions on Neural Networks and Learning Systems*, 35(5):6835–6849, 2022.

Jinyue Jiao, Zhiqiang Gong, and Ping Zhong. Triplet spectralwise transformer network for hyperspectral target detection. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–17, 2023.

Xudong Kang. Airport-beach-urban (ABU) datasets. `https://xudongkang.weebly.com/data-sets.html`, 2017.

Ryuichi Kiryo, Gang Niu, Marthinus C Du Plessis, and Masashi Sugiyama. Positive-unlabeled learning with non-negative risk estimator. *Advances in neural information processing systems*, 30, 2017.

Shawn Kraut, Louis L Scharf, and L Todd McWhorter. Adaptive subspace detectors. *IEEE Transactions on signal processing*, 49(1):1–16, 2001.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Haoyang Li, Xuejia Chen, Zhanchao Xu, Darian Li, Nicole Hu, Fei Teng, Yiming Li, Luyu Qiu, Chen Jason Zhang, Qing Li, et al. Exposing numeracy gaps: A benchmark to evaluate fundamental numerical abilities in large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 20004–20026, 2025a.

Jiayi Li, Xin Huang, and Lilin Tu. Whu-ohs: A benchmark dataset for large-scale hersepctral image classification. *International Journal of Applied Earth Observation and Geoinformation*, 113: 103022, 2022.

Jingtao Li, Yingyi Liu, Xinyu Wang, Yunning Peng, Chen Sun, Shaoyu Wang, Zhendong Sun, Tian Ke, Xiao Jiang, Tangwei Lu, Anran Zhao, and Yanfei Zhong. Hyperfree: A channel-adaptive and tuning-free foundation model for hyperspectral remote sensing imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23048–23058, June 2025b.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv*, 2026.

Wei Li and Qian Du. Collaborative representation for hyperspectral anomaly detection. *IEEE Transactions on geoscience and remote sensing*, 53(3):1463–1474, 2014.

Wei Li and Qian Du. Collaborative representation for hyperspectral anomaly detection. *IEEE Transactions on Geoscience and Remote Sensing*, 53(3):1463–1474, 2015. doi: 10.1109/TGRS.2014.2343955.

Xiang Li, Congcong Wen, Yuan Hu, Zhenghang Yuan, and Xiao Xiang Zhu. Vision-language models in remote sensing: Current progress and future trends. *IEEE Geoscience and Remote Sensing Magazine*, 12(2):32–66, 2024a.

Xuyang Li, Danfeng Hong, and Jocelyn Chanussot. S2mae: A spatial-spectral pretraining foundation model for spectral remote sensing data. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27696–27705. IEEE, 2024b.

Bing Liu, Xuchu Yu, Anzhu Yu, Pengqiang Zhang, Gang Wan, and Ruirui Wang. Deep few-shot learning for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 57(4):2290–2304, 2018.

Chengyang Liu, Aili Wang, Minhui Wang, Haibin Wu, Siqi Yan, and Lin Zhao. Cross-domain hyperspectral image classification combined sharpness-aware minimization with local-to-global feature enhancement. *Remote Sensing*, 18(5):740, 2026.

Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. Remoteclip: A vision language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–16, 2024.

Sicong Liu, Daniele Marinelli, Lorenzo Bruzzone, and Francesca Bovolo. A review of change detection in multitemporal hyperspectral images: Current techniques, applications, and challenges. *IEEE Geoscience and Remote Sensing Magazine*, 7(2):140–158, 2019.

Dimitris Manolakis, Eric Truslow, Michael Pieper, Thomas Cooley, and Michael Brueggeman. Detection algorithms in hyperspectral imaging systems: An overview of practical algorithms. *IEEE Signal Processing Magazine*, 31(1):24–33, 2013.

Larry R Medsker, Lakhmi Jain, et al. Recurrent neural networks. *Design and applications*, 5(64-67): 2, 2001.

Farid Melgani and Lorenzo Bruzzone. Classification of hyperspectral remote sensing images with support vector machines. *IEEE Transactions on Geoscience and Remote Sensing*, 42(8):1778–1790, 2004a.

Farid Melgani and Lorenzo Bruzzone. Classification of hyperspectral remote sensing images with support vector machines. *IEEE Transactions on geoscience and remote sensing*, 42(8):1778–1790, 2004b.

Sachit Menon and Carl Vondrick. Visual classification via description from large language models. *arXiv preprint arXiv:2210.07183*, 2022.

Shaheer Mohamed, Maryam Haghighat, Tharindu Fernando, Sridha Sridharan, Clinton Fookes, and Peyman Moghadam. Factoformer: Factorized hyperspectral transformers with self-supervised pretraining. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2023.

José Manuel Molero, Ester M. Garzon, Inmaculada Garcia, and Antonio Plaza. Analysis and optimizations of global and local versions of the rx algorithm for anomaly detection in hyperspectral data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 6(2): 801–814, 2013.

Allan Aasbjerg Nielsen. The regularized iteratively reweighted mad method for change detection in multi-and hyperspectral data. *IEEE Transactions on Image processing*, 16(2):463–478, 2007.

Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15645–15655. IEEE, 2023.

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*, pp. 8748–8763. PmLR, 2021.

Irving S Reed and Xiaoli Yu. Adaptive multiple-band cfar detection of an optical pattern with unknown spectral distribution. *IEEE transactions on acoustics, speech, and signal processing*, 38 (10):1760–1770, 1990.

Swalpa Kumar Roy, Gopal Krishna, Shiv Ram Dubey, and Bidyut B Chaudhuri. Hybridsn: Exploring 3-d–2-d cnn feature hierarchy for hyperspectral image classification. *IEEE geoscience and remote sensing letters*, 17(2):277–281, 2019.

Bernhard Schölkopf, Robert C Williamson, Alex Smola, John Shawe-Taylor, and John Platt. Support vector method for novelty detection. *Advances in neural information processing systems*, 12, 1999.

Dunbin Shen, Xiaorui Ma, Wenfeng Kong, Jianjun Liu, Jie Wang, and Hongyu Wang. Hyperspectral target detection based on interpretable representation network. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–16, 2023.

Kyle Stein, Andrew Arash Mahyari, Guillermo Francia, and Eman El-Sheikh. Visual adaptive prompting for compositional zero-shot learning. In *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 4196–4205. IEEE, 2025.

Le Sun, Guangrui Zhao, Yuhui Zheng, and Zebin Wu. Spectral–spatial feature tokenization transformer for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2022.

Lilin Tu, Jiayi Li, Xin Huang, Jianya Gong, Xing Xie, and Leiguang Wang. S 2 hm 2: A spectral–spatial hierarchical masked modeling framework for self-supervised feature learning and classification of large-scale hyperspectral images. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–19, 2024.

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.

Miguel A. Veganzones, Guillaume Tochon, Mauro Dalla-Mura, Antonio J. Plaza, and Jocelyn Chanussot. Hyperspectral image segmentation using a new spectral unmixing-based binary partition tree representation. *IEEE Transactions on Image Processing*, 23(8):3574–3589, 2014.

Di Wang, Meiqi Hu, Yao Jin, Yuchun Miao, Jiaqi Yang, Yichu Xu, Xiaolei Qin, Jiaqi Ma, Lingyu Sun, Chenxing Li, et al. Hypersigma: Hyperspectral intelligence comprehension foundation model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(8):6427–6444, 2025.

Shaoyu Wang, Xinyu Wang, Liangpei Zhang, and Yanfei Zhong. Auto-ad: Autonomous hyperspectral anomaly detection network based on fully convolutional autoencoder. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2022a. doi: 10.1109/TGRS.2021.3057721.

Yanheng Wang, Danfeng Hong, Jianjun Sha, Lianru Gao, Lian Liu, Yonggang Zhang, and Xianhui Rong. Spectral–spatial–temporal transformers for hyperspectral image change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2022b.

Yue Wang, Ming Wen, Hailiang Zhang, Jinyu Sun, Qiong Yang, Zhimin Zhang, and Hongmei Lu. Hsimae: A unified masked autoencoder with large-scale pretraining for hyperspectral image classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17:14064–14079, 2024.

Xiaodong Wei, Ruizhe Liu, Jiyuan Li, and Hongmin Gao. S2tmn: A spectral-spatial-temporal mamba network for hyperspectral image change detection. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2026.

Wen Wu, Ziyang Zhang, Liwei Liu, Xuenan Xu, Jimin Zhuang, Ke Fan, Qitan Lv, Junlin Liu, Chen Zhang, Zheqi Yuan, et al. Scits: Scientific time series understanding and generation with llms. In *International Conference on Learning Representations*, volume 2026, pp. 114390–114431, 2026.

Yang Xu, Zebin Wu, Jun Li, Antonio Plaza, and Zhihui Wei. Anomaly detection in hyperspectral images based on low-rank and sparse representation. *IEEE Transactions on Geoscience and Remote Sensing*, 54(4):1990–2000, 2015.

Yichu Xu, Lefei Zhang, Bo Du, and Liangpei Zhang. Hyperspectral anomaly detection based on machine learning: An overview. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:3351–3364, 2022.

Chuguang Zeng, Jingtao Li, Yinhe Liu, and Yanfei Zhong. Hyperimagenet: A large-scale high-spatial resolution hyperspectral imagery classification benchmark. *arXiv preprint arXiv:2607.21050*, 2026.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11941–11952. IEEE, 2023.

Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*, 2021.

Renrui Zhang, Xiangfei Hu, Bohao Li, Siyuan Huang, Hanqiu Deng, Yu Qiao, Peng Gao, and Hongsheng Li. Prompt, generate, then cache: Cascade of foundation models makes strong few-shot learners. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15211–15222. IEEE, 2023.

Zhimin Zhang, Chengzhen Ma, Xiao Teng, Huansheng Ning, Lingfeng Mao, and Dan Zhang. Hyperspectral anomaly detection based on spatial-spectral feature fusion autoencoder. *Neurocomputing*, pp. 133162, 2026.

Zilun Zhang, Tiancheng Zhao, Yulong Guo, and Jianwei Yin. Rs5m and georsclip: A large-scale vision-language dataset and a large vision-language model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–23, 2024.

Hengwei Zhao, Xinyu Wang, Jingtao Li, and Yanfei Zhong. Class prior-free positive-unlabeled learning with taylor variational loss for hyperspectral remote sensing imagery. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16781–16790. IEEE, 2023a.

Hengwei Zhao, Yanfei Zhong, Xinyu Wang, and Hong Shu. One-class risk estimation for one-class hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–17, 2023b.

Yanfei Zhong, Xin Hu, Chang Luo, Xinyu Wang, Ji Zhao, and Liangpei Zhang. WHU-Hi: UAV-borne hyperspectral with high spatial resolution (H2) benchmark datasets and classifier for precise crop identification based on deep convolutional neural network with CRF. *Remote Sensing of Environment*, 250:112012, 2020. doi: 10.1016/j.rse.2020.112012.

Zilong Zhong, Jonathan Li, Zhiming Luo, and Michael Chapman. Spectral–spatial residual network for hyperspectral image classification: A 3-d deep learning framework. *IEEE transactions on geoscience and remote sensing*, 56(2):847–858, 2017.

Lingxuan Zhu, Jiaji Wu, Wang Biao, Yi Liao, and Dandan Gu. Spectralmae: Spectral masked autoencoder for hyperspectral remote sensing image reconstruction. *Sensors*, 23(7):3728, 2023.

APPENDIX

A Downstream Task Applications **16**

B VLM backbones **17**

C Extended Experimental Results and Visualizations **17**

C.1 Extended Experimental Results of Classification 18

C.2 Extended Anomaly Detection and RX Enhancement 18

C.3 Additional Qualitative Visualizations 19

D Limitations and Future Work **20**

E Dataset Information and Evaluation Metrics **20**

E.1 Hyperspectral Image Datasets 20

E.2 Evaluation Metrics 22

A DOWNSTREAM TASK APPLICATIONS

HyperMAG extracts unified, domain-agnostic multimodal features that directly support diverse hyperspectral tasks, including few-shot classification, one-class classification, target detection, anomaly detection, and change detection. To mitigate single-scale segmentation bias, final pixel-level predictions are generated by aggregating region-level outputs across multiple scales $c \in \{c_1, \ldots, c_C\}$ via majority voting for discrete labels and average pooling for continuous scores.

Few-Shot Classification. Given a few labeled anchor pixels, we average their region-level features to construct class prototypes $\mathbf{z}^{(i)}$ for each class $i \in \{1, \ldots, K\}$. For a query region r with feature $\mathbf{z}^{(r)}$, its initial class prediction $\hat{y}_r$ is assigned via maximum cosine similarity to the prototypes

$$S_{r,i} = \frac{\mathbf{z}^{(r)\top}\mathbf{z}^{(i)}}{\|\mathbf{z}^{(r)}\|_2 \|\mathbf{z}^{(i)}\|_2}, \quad \hat{y}_r = \arg\max_{i \in \{1,\ldots,K\}} S_{r,i}. \tag{8}$$

To mitigate single-scale segmentation bias, the final classification label for pixel e is obtained by ensembling predictions across all C scales via majority voting

$$\hat{y}_e = \arg\max_{k \in \{1,\ldots,K\}} \sum_{c=1}^{C} \mathbb{I}\big[\hat{y}_e^{(c)} = k\big], \tag{9}$$

where $\hat{y}_e^{(c)}$ denotes the predicted region label containing pixel e at scale c, and $\mathbb{I}[\cdot]$ is the indicator function.

One-Class Classification. For one-class classification, we aggregate the features of the available positive reference regions $\mathcal{R}_{\text{pos}}$ to form a normalized target category prototype $\mathbf{p}$

$$\mathbf{p} = \frac{1}{|\mathcal{R}_{\text{pos}}|} \sum_{r \in \mathcal{R}_{\text{pos}}} \mathbf{z}^{(r)}, \quad \mathbf{p} \leftarrow \frac{\mathbf{p}}{\|\mathbf{p}\|}. \tag{10}$$

The likelihood of a query region belonging to the target class is then determined by computing the cosine similarity between the query feature and the positive prototype.

Target Detection. Target detection aims to locate targets that match a specific reference spectrum $\bar{\mathbf{x}}_t$ and its corresponding multimodal feature $\mathbf{z}^{(t)}$. We compute a hybrid detection score $\hat{s}_r$ by fusing the semantic similarity in the VLM feature space with the physical Spectral Angle Mapper (SAM) in the original spectral space

$$\hat{s}_r = \alpha \cdot \frac{\langle \mathbf{z}^{(r)}, \mathbf{z}^{(t)} \rangle}{\|\mathbf{z}^{(r)}\| \cdot \|\mathbf{z}^{(t)}\|} + (1-\alpha) \cdot \left(1 - \frac{1}{\pi} \arccos\left(\frac{\langle \bar{\mathbf{x}}_r, \bar{\mathbf{x}}_t \rangle}{\|\bar{\mathbf{x}}_r\| \cdot \|\bar{\mathbf{x}}_t\|}\right)\right), \tag{11}$$

where $\alpha \in [0,1]$ is a hyperparameter balancing the semantic and physical metrics.

Change Detection. Given co-registered bi-temporal hyperspectral images $\mathbf{X}_1, \mathbf{X}_2 \in \mathbb{R}^{H \times W \times B}$, we first construct the absolute difference cube $\mathbf{X}_\Delta = |\mathbf{X}_2 - \mathbf{X}_1|$. Multi-scale region difference spectra $\bar{\mathbf{x}}_r$ are extracted from $\mathbf{X}_\Delta$ and transformed into unified multimodal features $\mathbf{z}^{(r)}$ via our multimodal translation module. Let $\mathcal{R}_0$ denote the set of unchanged background reference regions, and $\mathbf{p}$ be their normalized feature centroid

$$\mathbf{p} = \frac{\sum_{r \in \mathcal{R}_0} \mathbf{z}^{(r)}}{\left\| \sum_{r \in \mathcal{R}_0} \mathbf{z}^{(r)} \right\|_2}. \tag{12}$$

The change intensity score a_r for a query region r is evaluated via its cosine distance to the background centroid $\mathbf{p}$

$$a_r = 1 - \frac{\mathbf{z}^{(r)\top} \mathbf{p}}{\|\mathbf{z}^{(r)}\|_2}. \tag{13}$$

Finally, the robust pixel-level change map is generated by average-pooling the region change scores a_e of each pixel across all C scales: $a_e = \frac{1}{C} \sum_{c=1}^{C} a_e^{(c)}$.

Anomaly Detection. For anomaly detection, we adapt the classic Reed-Xiaoli (RX) detector (Reed & Yu, 1990) into our VLM feature space. Given the region features $\{\mathbf{z}^{(r)}\}_{r=1}^{R}$, we estimate the background mean $\boldsymbol{\mu}$ and the regularized covariance matrix $\boldsymbol{\Sigma}$

$$\boldsymbol{\mu} = \frac{1}{R} \sum_{r=1}^{R} \mathbf{z}^{(r)}, \quad \boldsymbol{\Sigma} = \frac{1}{R-1} \sum_{r=1}^{R} \big(\mathbf{z}^{(r)} - \boldsymbol{\mu}\big)\big(\mathbf{z}^{(r)} - \boldsymbol{\mu}\big)^\top + \gamma \mathbf{I}, \tag{14}$$

where $\gamma \mathbf{I}$ represents a small identity regularization term to ensure numerical stability. The region-level VLM anomaly score a_r^{VLM} is then computed via the Mahalanobis distance

$$a_r^{\mathrm{VLM}} = \big(\mathbf{z}^{(r)} - \boldsymbol{\mu}\big)^\top \boldsymbol{\Sigma}^{-1} \big(\mathbf{z}^{(r)} - \boldsymbol{\mu}\big). \tag{15}$$

To exploit both high-level semantic distinctiveness and low-level physical variations, the final anomaly score a_r fuses a_r^{VLM} with the standard RX score a_r^{SPEC} derived from the original spectral space

$$a_r = \frac{1}{2} \mathcal{N}\big(a_r^{\mathrm{VLM}}\big) + \frac{1}{2} \mathcal{N}\big(a_r^{\mathrm{SPEC}}\big), \tag{16}$$

where $\mathcal{N}(\cdot)$ denotes spatial min-max normalization. For distinction, the baseline variant relying purely on VLM features (a_r^{VLM}) is denoted as **HyperMAG**, whereas the fused score (a_r) is designated as **HyperMAG***. As with other tasks, the pixel-level anomaly map a_e is obtained via multi-scale average pooling: $a_e = \frac{1}{C} \sum_{c=1}^{C} a_e^{(c)}$.

B VLM BACKBONES

This paper mainly employs four representative pre-trained VLM backbones in HyperMAG, such as CLIP ViT-B/16 (Radford et al., 2021), SigLIP 2 SoViT-400M (Zhai et al., 2023), Qwen3-VL-Embedding-2B (Li et al., 2026), and Qwen3-VL-Embedding-8B (Li et al., 2026). Detailed architectural specifications are summarized in Table 6.

C EXTENDED EXPERIMENTAL RESULTS AND VISUALIZATIONS

To complement the multi-task evaluations in the main text, we provide more comprehensive quantitative results on extended datasets, further algorithmic variants, and qualitative visualizations across various downstream tasks.

Table 6: Detial information of pre-trained Vision-Language Models (VLMs).

VLM	Developer	Params	Embedding Dim.	Context / Input Size	Encoder Architecture
CLIP ViT-B/16	OpenAI	∼150M	512	Text: 77 tokens / Image: 224×224	Dual-tower (Separate text & image)
SigLIP 2 SoViT-400M	Google	∼400M (Vision)	1152	Text: 256 tokens / Image: 384×384	Dual-tower (Separate text & image)
Qwen3-VL-Embedding-2B	Alibaba	∼2B	2048	8192 tokens	Joint Multimodal (Unified input)
Qwen3-VL-Embedding-8B	Alibaba	∼8B	4096	8192 tokens	Joint Multimodal (Unified input)

Table 7: Quantitative comparison of HyperMAG with state-of-the-art hyperspectral classification methods. **Bold** and underlined values denote the best and second-best performance, respectively.

Method	Type	Indian Pines			PaviaU			HongHu			LongKou			HanChuan		
		OA	AA	κ	OA	AA	κ	OA	AA	κ	OA	AA	κ	OA	AA	κ
SVM	Traditional	40.89	54.82	0.35	55.83	70.97	0.48	51.05	50.02	0.44	79.07	70.35	0.74	54.51	46.36	0.48
HybridSN	CNN	55.30	67.80	0.50	63.48	61.59	0.52	55.70	34.24	0.47	81.77	64.00	0.77	42.67	26.48	0.28
SpectralFormer	Transformer	58.10	71.50	0.53	62.23	66.24	0.52	43.64	35.31	0.36	82.23	51.91	0.77	43.26	26.46	0.32
SSFTT	Transformer	60.50	73.20	0.56	69.72	71.92	0.61	61.35	49.21	0.53	88.43	78.88	0.85	59.98	46.23	0.53
HyperMAG	VLM	**73.08**	**77.94**	**0.70**	**84.54**	**87.03**	**0.80**	**76.78**	**71.48**	**0.71**	**89.03**	**87.05**	**0.86**	**71.80**	**67.53**	**0.67**

C.1 Extended Experimental Results of Classification

Comparison with Task-Specific Few-Shot Classification Methods. In addition to the LongKou and HanChuan datasets presented in the main text, we extensively evaluate the 5-shot classification performance of HyperMAG on more diverse datasets, including Indian Pines, Pavia University, and WHU-Hi-HongHu. The detailed quantitative results are summarized in Table 7. Table 7 presents the quantitative 5-shot classification comparison across Overall Accuracy (OA), Average Accuracy (AA), and Kappa coefficient (κ). HyperMAG consistently outperforms all competitive baselines across all five benchmark scenes by a substantial margin. Specifically, on challenging datasets with high spectral similarity like Indian Pines and HanChuan. Notably, unlike supervised or domain-specifically fine-tuned baselines, HyperMAG requires zero gradient updates. These results confirm that translating raw spectral curves into VLM-native textual and visual modalities enables frozen off-the-shelf VLMs to extract domain-agnostic, highly discriminative representations.

Few-Shot Scalability and Stability. To investigate the scalability with varying labeled data, Table 8 reports performance under 5-shot, 10-shot, and 20-shot settings, with mean accuracies and standard deviations computed over 10 independent runs using random seeds. As the shot budget expands from 5 to 20, HyperMAG exhibits monotonic performance growth alongside reduced variance across all scenes. On PaviaU, for instance, OA scales from 84.54% ($\pm$6.94%) at 5-shot to 94.74% ($\pm$1.59%) at 20-shot. This trend highlights that our multimodal prototype formulation effectively constructs stable class centroids even with minimal labeled anchors.

C.2 Extended Anomaly Detection and RX Enhancement

For the unsupervised hyperspectral anomaly detection (HAD) task, we extend our evaluation to all 13 sub-scenes of the ABU dataset and the HYDICE dataset. Furthermore, to fully leverage both the semantic representations of the VLM and the physical properties of the raw data, we evaluate two variants of our approach:

- **HyperMAG**: Applying the RX operator directly within the semantic feature space extracted by the frozen VLM.
- **HyperMAG***: Utilizing the raw spectral information to enhance the VLM feature-based RX detection results.

The comprehensive AUROC results are reported in Table 9. Even when applied directly in the feature space (HyperMAG), our method generally outperforms the HyperFree baseline. With the raw spectral enhancement (HyperMAG*), the model achieves highly competitive detection accuracy, matching or exceeding specialized HAD models across multiple challenging scenes. Notably, these highly competitive results are achieved in a completely training-free manner, requiring zero

Table 8: Quantitative few-shot classification results of HyperMAG on five benchmark datasets. Performance variations are highlighted with light, medium, and dark red backgrounds, respectively.

No. of samples	Dataset	OA (%)	AA (%)	Kappa (%)
5-shot	PaviaU	84.54 ± 6.94	87.03 ± 1.90	80.33 ± 8.21
	IndianPines	73.08 ± 4.37	77.94 ± 2.60	69.74 ± 4.75
	HanChuan	71.80 ± 2.46	67.53 ± 2.56	67.28 ± 2.69
	HongHu	76.78 ± 2.49	71.48 ± 1.45	71.43 ± 2.71
	LongKou	89.03 ± 3.04	87.05 ± 3.08	85.95 ± 3.79
10-shot	PaviaU	90.61 ± 4.06	90.10 ± 2.28	87.76 ± 4.87
	IndianPines	81.13 ± 3.40	79.55 ± 4.24	78.66 ± 3.81
	HanChuan	80.50 ± 1.91	77.42 ± 1.92	77.28 ± 2.20
	HongHu	81.19 ± 2.23	75.79 ± 2.15	76.71 ± 2.52
	LongKou	94.71 ± 1.40	93.30 ± 1.07	93.14 ± 1.78
20-shot	PaviaU	94.74 ± 1.59	93.92 ± 1.08	93.05 ± 2.02
	IndianPines	84.41 ± 1.99	76.52 ± 3.29	82.33 ± 2.18
	HanChuan	86.08 ± 1.01	84.16 ± 1.50	83.77 ± 1.18
	HongHu	85.06 ± 1.04	77.49 ± 1.93	81.25 ± 1.24
	LongKou	96.67 ± 0.53	95.86 ± 0.65	95.66 ± 0.69

Table 9: Quantitative comparison of HyperMAG against state-of-the-art hyperspectral anomaly detection methods. **Bold** and underlined values denote the best and second-best performance, respectively.

	Task-specific HAD Methods				Foundation-model-based Methods			
Dataset	RX	CRD	Auto-AD	GT-HAD	HyperFree	HyperSIGMA	HyperMAG	HyperMAG*
Airport-1	82.21	85.91	93.54	**96.26**	75.44	61.07	77.27	87.49
Airport-2	84.04	78.05	92.04	**98.63**	56.72	69.94	92.29	94.78
Airport-3	92.88	87.47	87.06	**96.20**	60.85	91.30	81.56	93.94
Airport-4	95.26	71.51	97.88	99.13	69.35	**99.49**	86.64	95.15
Airport-Mean	88.60	80.73	92.63	**97.56**	65.59	80.45	84.44	92.84
Beach-1	98.07	97.37	96.42	**99.92**	97.27	96.02	98.03	97.18
Beach-2	91.06	83.48	93.03	91.57	**93.14**	90.64	81.46	90.72
Beach-3	**99.98**	99.94	99.93	**99.98**	49.40	92.23	93.17	99.79
Beach-4	95.38	94.62	98.11	**98.70**	92.28	91.34	90.04	97.35
Beach-Mean	96.12	93.85	96.87	**97.54**	83.02	92.56	90.68	96.26
Urban-1	99.07	98.23	99.28	**99.72**	76.06	97.97	96.16	99.27
Urban-2	99.46	96.81	99.00	99.56	84.87	99.71	96.29	**99.81**
Urban-3	95.13	84.19	**98.15**	97.37	91.07	79.77	72.57	94.26
Urban-4	98.87	90.51	98.86	98.93	60.67	**99.60**	94.34	98.47
Urban-5	**96.92**	92.92	86.08	92.26	73.58	96.61	85.97	95.39
Urban-Mean	**97.89**	92.53	96.27	97.57	77.25	94.73	89.07	97.44
HYDICE	97.63	81.72	98.60	90.96	80.49	92.16	90.64	**99.46**

fine-tuning or task-specific optimization while remaining fully comparable to existing specialized methods and trainable foundation models.

C.3 Additional Qualitative Visualizations

To provide a more intuitive understanding of the multi-task performance of HyperMAG, we supply additional visual prediction maps for the multiple tasks. The qualitative visualizations across all downstream tasks—including the few-shot classification maps (Fig. 8), one-class classification results (Fig. 11), anomaly detection probability maps (Fig. 9), change detection maps (Fig. 10), further support our quantitative evaluations. Specifically, the visual outputs for few-shot classification and change detection demonstrate that HyperMAG maintains strong inter-class discriminability and precise boundary delineation under extremely limited supervision. Meanwhile, the response maps for one-class and anomaly detection confirm that after spectral-to-text and spectral-to-image translation, intra-class features preserve high spatial compactness and target saliency in the joint visual-textual feature space. Overall, these qualitative comparisons intuitively reinforce our numerical findings,

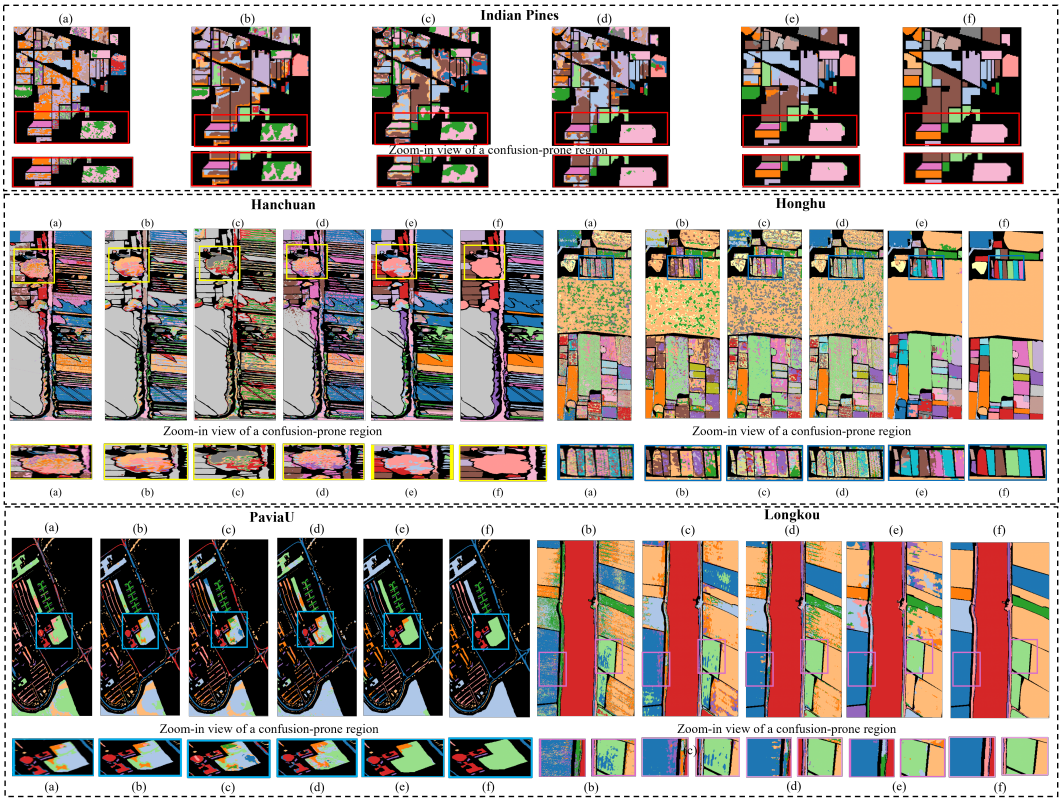

Figure 8: Visual comparison of 5-shot classification results on the Indian Pines, HanChuan, HongHu, PaviaU, and LongKou datasets. Columns (a)-(f) show the classification maps of SVM, HybridSN, SpectralFormer, SSFTT, HyperMAG, and Ground Truth, respectively. Different colors denote semantic classes, and black denotes unlabeled/background regions. The enlarged regions highlight local differences around complex urban structures.

validating that HyperMAG achieves remarkable multi-task generalization capability and robustness in a completely training-free and fine-tuning-free manner.

D Limitations and Future Work

HyperMAG has limitations despite its training-free efficiency. First, multi-scale superpixels and dual-branch encoding introduce higher inference latency than task-specific baselines. Second, processing regions independently overlooks 3D spatial-spectral continuity in HSI cubes. Future work will explore lightweight learnable projectors or adaptive tokenization to lower latency while maintaining the training-free backbone. Additionally, this cross-modal strategy can scale to other non-natural scientific modalities, such as sequential data, expanding VLMs beyond standard vision and text modalities.

E Dataset Information and Evaluation Metrics

E.1 Hyperspectral Image Datasets

Below, we briefly introduce the datasets used in our experiments, with their detailed specifications summarized in Table 10.

WHU-Hi Dataset. The WHU-Hi dataset (Zhong et al., 2020) consists of three UAV-borne hyperspectral scenes (LongKou, HanChuan, and HongHu) acquired over agricultural areas in China. Featuring diverse crop types and complex spatial distributions.

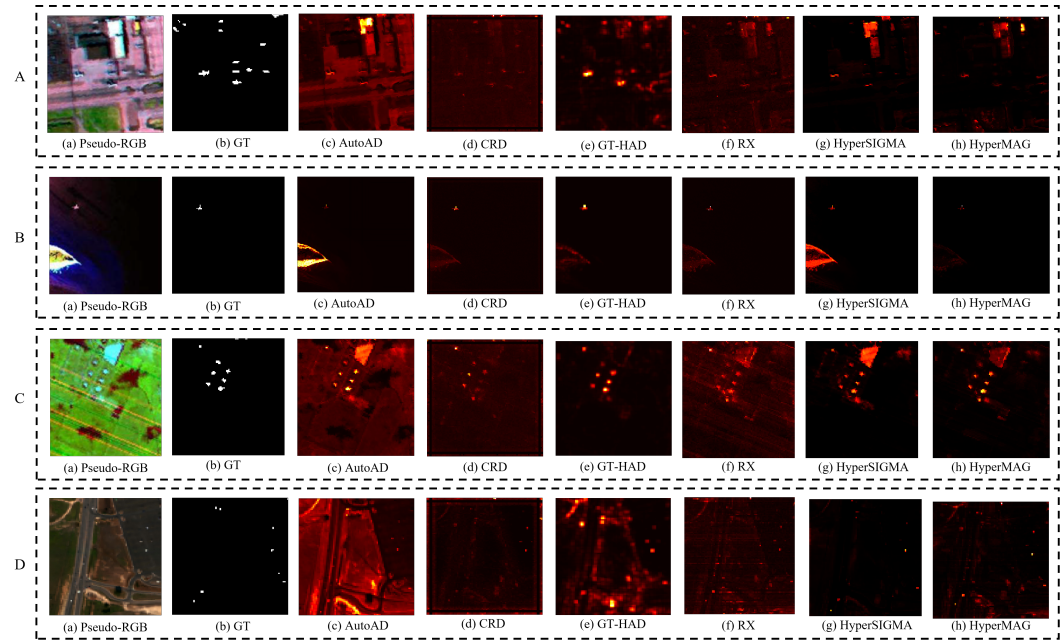

Figure 9: Visualization of anomaly detection results on four representative scenes. Rows A-D correspond to scenes from the Airport, Beach, and Urban subsets. Columns (a)-(h) show (a) the pseudo-RGB image, (b) ground truth, and the anomaly maps produced by (c) Auto-AD, (d) CRD, (e) GT-HAD, (f) RX, (g) HyperSIGMA, and (h) HyperMAG (Ours), respectively.

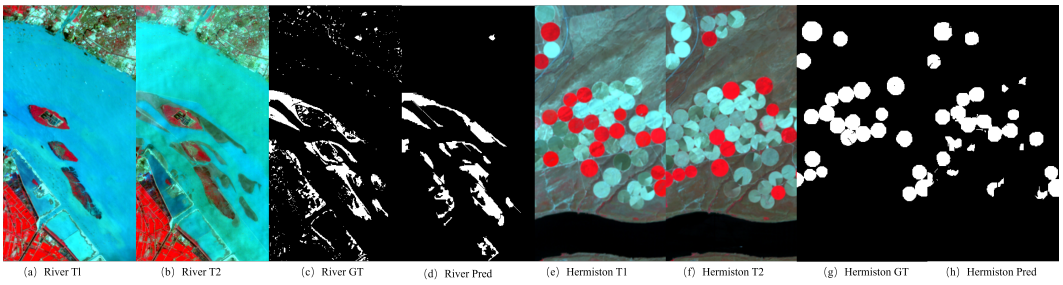

Figure 10: Visual comparison of change detection results on the River and Hermiston datasets. (a)(b) and (e)(f) present the bi-temporal pseudo-RGB image pairs, (c) and (g) show the corresponding ground truth maps, while (d) and (h) illustrate the change detection maps generated by the proposed HyperMAG.

Indian Pines Dataset. Acquired by the AVIRIS sensor over Indiana, USA, the Indian Pines dataset (Baumgardner et al., 2015) comprises 145×145 pixels with a 20 m spatial resolution. After removing noisy channels, it retains 200 valid spectral bands and provides annotations for 16 agricultural and vegetation categories.

Pavia University Dataset. Collected by the ROSIS sensor over Pavia, Italy, this dataset (Huang & Zhang, 2009) contains 610×340 pixels at a 1.3 m spatial resolution. It features 103 effective spectral bands and includes annotations for nine typical urban land-cover classes.

XiongAn Dataset. Acquired over the Xiong'an New Area, China, the XiongAn dataset (Cen et al., 2020) has a cropped size of 3750×1580 pixels with a high spatial resolution of 0.5 m. It covers 250 spectral bands and provides annotations for 20 land-cover categories, primarily cash crops.

ABU Dataset. Cropped from AVIRIS imagery, the ABU dataset (Kang, 2017) contains 13 sub-scenes categorized into three background types: Airport (4), Beach (4), and Urban (5). These diverse

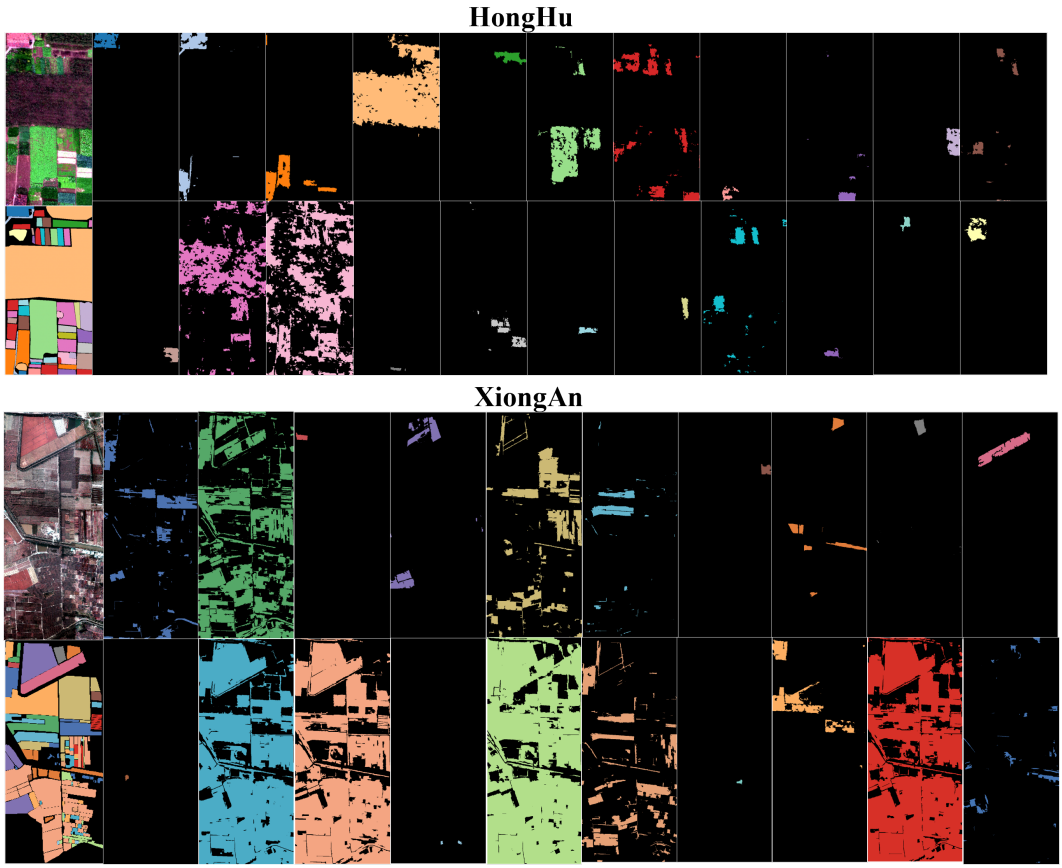

Figure 11: Visualization of one-class classification results on the HongHu (22 classes) and XiongAn (20 classes) datasets. The first column shows the pseudo-RGB images and ground truth maps of each dataset. The remaining columns present the one-class classification results for each class in these two datasets.

scenes are widely utilized to evaluate anomaly detection performance under varying background complexities.

HYDICE Dataset. Acquired by the HYDICE sensor, this dataset (Li & Du, 2014) features an urban scene with a 1 m spatial resolution. In our experiments, an 80×100 cropped region is utilized, where vehicles of varying sizes (21 pixels in total) are defined as anomalies.

Hermiston Dataset. The Hermiston dataset (Hu et al., 2024) is a bitemporal hyperspectral dataset acquired by the Hyperion sensor in 2004 and 2007. It contains 307×241 pixels and 154 bands, with changes mainly attributed to crop growth and agricultural transitions.

River Dataset. Acquired in May and December 2013 over Jiangsu, China, the River dataset (Hu et al., 2024) provides bitemporal scenes of size 463×241 pixels with 198 valid bands. The primary change lies in the disappearance of substances within the river.

E.2 Evaluation Metrics

To comprehensively evaluate model performance across diverse hyperspectral interpretation tasks, we adopt specific evaluation metrics tailored to each application.

Classification Metrics. For hyperspectral image classification (Liu et al., 2018), let $M \in \mathbb{R}^{K \times K}$ be the confusion matrix for K classes, where M_{ij} denotes the number of class i samples predicted as class j. Overall Accuracy (OA) measures the global proportion of correctly classified samples,

Table 10: Detailed Description of Hyperspectral Image Datasets.

Classification	Class Number	Resolution	Size
WHU-Hi-LongKou	9	0.463m	$550 \times 400 \times 270$
WHU-Hi-HanChuan	16	0.019m	$1217 \times 303 \times 274$
WHU-Hi-HongHu	22	0.043m	$940 \times 475 \times 270$
Indian Pines	16	20m	$145 \times 145 \times 200$
PaviaU	9	1.3m	$610 \times 340 \times 103$
XiongAN	20	0.5m	$3750 \times 1580 \times 250$
Anomaly Detection	Captured Place	Resolution	Size
Airport-1	Los Angeles	7.1m	$100 \times 100 \times 205$
Airport-2	Los Angeles	7.1m	$100 \times 100 \times 205$
Airport-3	Los Angeles	7.1m	$100 \times 100 \times 205$
Airport-4	Gulfport	3.4m	$100 \times 100 \times 191$
Beach-1	Cat Island	17.2m	$150 \times 150 \times 188$
Beach-2	San Diego	7.5m	$100 \times 100 \times 193$
Beach-3	Bay Champagne	4.4m	$100 \times 100 \times 188$
Beach-4	Pavia	1.3m	$150 \times 150 \times 102$
Urban-1	Texas Coast	17.2m	$100 \times 100 \times 204$
Urban-2	Texas Coast	17.2m	$100 \times 100 \times 207$
Urban-3	Gainesville	3.5m	$100 \times 100 \times 191$
Urban-4	Los Angeles	7.1m	$100 \times 100 \times 205$
Urban-5	Los Angeles	7.1m	$100 \times 100 \times 205$
HYDICE-Urban	Michigan	2m	$100 \times 80 \times 162$
Change Detection	Captured Time	Resolution	Size
River	2013.5.3–2013.12.31	30m	$463 \times 241 \times 198$
Hermiston	2004.5.1–2007.5.8	30m	$307 \times 241 \times 154$

$\mathrm{OA} = \frac{\sum_{i=1}^{K} M_{ii}}{\sum_{i=1}^{K}\sum_{j=1}^{K} M_{ij}}$. To mitigate the bias of OA toward dominant classes under category imbalance, Average Accuracy (AA) computes the unweighted arithmetic mean of per-class accuracy, $\mathrm{AA} = \frac{1}{K}\sum_{i=1}^{K} \frac{M_{ii}}{\sum_{j=1}^{K} M_{ij}}$. Furthermore, the Kappa coefficient (κ) evaluates the statistical agreement between predictions and ground-truth labels while accounting for chance agreement.

Target and Anomaly Detection Metrics. For hyperspectral target and anomaly detection (Chen et al., 2023; Xu et al., 2022), we treat target or anomalous pixels as the positive class and background as the negative class. The primary evaluation metric is the Area Under the Receiver Operating Characteristic curve (AUROC). The ROC curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) across varying decision thresholds, where $\mathrm{TPR} = TP/(TP + FN)$ and $\mathrm{FPR} = FP/(FP + TN)$. AUROC values closer to 1 indicate superior separability between targets/anomalies and the background.

One-Class Classification Metrics. In one-class classification (Zhao et al., 2023b), target samples are defined as the positive class. We evaluate performance using Precision, $\mathrm{Precision} = TP/(TP + FP)$, Recall, $\mathrm{Recall} = TP/(TP+FN)$, and the F1-score, $\mathrm{F1} = 2\cdot\mathrm{Precision}\cdot\mathrm{Recall}/(\mathrm{Precision}+\mathrm{Recall})$. For multi-class scenarios, each class is evaluated individually, and we report the average F1-score and Recall values.

Change Detection Metrics. Change detection is formulated as a pixel-wise binary classification task where changed pixels are positive (Liu et al., 2019). Due to the severe class imbalance typically dominated by unchanged background pixels, we employ Intersection over Union (IoU) and the F1-score rather than global accuracy. IoU measures the spatial overlap between predicted and actual changed regions: $\mathrm{IoU} = TP/(TP + FP + FN)$. Both metrics strictly penalize false alarms and missed detections, with higher values signifying better consistency with the ground truth.