Learning Hyperspectral Images with Curated Text Prompts for Efficient Multimodal Alignment

Abhiroop Chatterjee, Susmita Ghosh
Jadavpur University
{abhiroopchat1998, susmitaghoshju}@gmail.com

Abstract

As data requirements continue to grow, efficient learning increasingly depends on the curation and distillation of high-value data rather than brute-force scaling of model sizes. In the case of a hyperspectral image (HSI), the challenge is amplified by the high-dimensional 3D voxel structure, where each spatial location is associated with hundreds of contiguous spectral channels. While vision and language models have been optimized effectively for natural image or text tasks, their cross-modal alignment in the hyperspectral domain remains an open and underexplored problem. In this article, we make an attempt to optimize a Vision–Language Model (VLM) for hyperspectral scene understanding by exploiting a CLIP-style contrastive training framework. Our framework maps voxel-level embeddings from a vision backbone onto the latent space of a frozen large embedding model (LEM), where a trainable probe aligns vision features with the model's textual token representations. The two modalities are aligned via a contrastive loss restricted to a curated set of hard (closest wrong classes) and semi-hard (random distractors) negatives, along with positive pairs. To further enhance alignment, descriptive prompts that encode class semantics are introduced and act as structured anchors for the HSI embeddings. It is seen that the proposed method updates only ***0.07%*** *of the total parameters, yet yields state-of-the-art performance. For example, on Indian Pines (IP) the model produces better results over unimodal and multimodal baselines by +0.92 Overall Accuracy (OA) and +1.60 Kappa (κ), while on Pavia University (PU) data it provides gains of +0.69 OA and +0.90 κ. Moreover, this is achieved with the set of parameters, nearly* ***50***$\times$ *smaller than DCTN and* ***90***$\times$ *smaller than SS-TMNet.*

1. Introduction

Hyperspectral imaging (HSI) [16, 25] captures rich spatial–spectral information across hundreds of contiguous narrow bands, and facilitates fine-grained analysis of material properties and scene characteristics. Unlike RGB or multispectral imagery, which provide only a handful of broad channels, HSI encodes detailed spectral signatures that can be used to distinguish between objects that appear visually identical. This capability has made hyperspectral methods indispensable in domains [14, 18, 20, 37] such as *remote sensing*, *environmental surveillance*, *defense and security*, and *biomedical imaging*.

At the same time, this high-dimensional 3D data structure introduces fundamental challenges for *representation learning* [6]. Spectral redundancy, strong inter-band correlation, and the curse of dimensionality complicate feature extraction, while limited labeled datasets amplify the risk of overfitting. Models must also preserve the **spatial context** of scenes, critical for interpreting patterns such as vegetation distribution or tumor boundaries, and at the same time, exploit the fine-grained spectral features that distinguish materials. Moreover, HSI tasks differ from conventional vision problems in several respects. We are listing some of them below:

1. **Spectral precision** – Success often depends on identifying minute spectral differences invisible to RGB-based systems.
2. **Data limitations** – High acquisition cost and complex sensor setups result in smaller, domain-specific datasets compared to large-scale benchmarks such as ImageNet.
3. **Evaluation protocol – Labeling hyperspectral pixels is extremely difficult and expensive**, the evaluation setup [39] is often **reversed compared to standard vision tasks**. Instead of training on large annotated sets and testing on small subsets, HSI models are commonly trained with only **10% labeled training data** and evaluated on the remaining **90% test data**. This type of protocol reflects the practical reality of scarce supervision and shows the need for efficient learning methods under limited label scenarios.

These differences raise the need to build newer methods to progress hyperspectral analysis. Recent advances have explored diverse vision architectures [5, 8, 11, 15, 33] to

address these challenges. We highlight the **state-of-the-art methods** [13, 28, 29, 31, 32, 35, 39] along with **established world models** [8, 9, 17, 26] below:

Literature Review. Early HSI models relied on convolutional approaches. **2D-CNN** [31] extracts spatial features via stacked 2-D convolutions, while **3D-CNN** [35] jointly models spatial–spectral information. Hybrid models like **HybridSN** [28] combine 3D and 2D convolutions, and capture both spatial–spectral and spatial features. Transformer-based methods, including **ViT** [8, 10], **SSFTT** [32], and **morphFormer** [29], exploit attention to model long-range dependencies. Other architectures such as **HiT** [36] and **SS-TMNet** [13] use spectral A3D convolutions and multiscale spatial–spectral attention. On the other hand, dual-branch networks like **DCTN** [39], combine CNNs for local features with transformers for global spectral modeling to achieve state-of-the-art performance. Contrastive vision–language frameworks [19, 38] such as **CLIP** [26] provide transferable cross-modal embeddings, but are designed for natural 2D images and overlook volumetric spatial-spectral structures in HSI. Several remote sensing applications often require understanding complex 3D patterns, and this motivates us to explore the multimodal alignment [26] in the hyperspectral domain.

Here, we present a *Vision Language Model* (Figure 1a), for *hyperspectral scene understanding*, that aligns a ViT backbone with a frozen large embedding model (LEM) [7, 22, 34] through contrastive learning. To boost discriminability, we employ *descriptive prompts* encoding class semantics and *informative negatives* to counter hard distractors. It is seen that training around 0.07% of parameters, our method achieves SOTA on various HSI benchmarks. A parameter efficiency snapshot is given in Figure 1b.

The article is organized as follows: Section 2 details the proposed methodology, while Section 3 describes the experimental setup. Section 4 presents the results with analysis, and Section 5 concludes with an outlook and future directions.

2. Methodology

Let $\mathcal{X} \subset \mathbb{R}^{H\times W\times D}$ denote hyperspectral patches and $\mathcal{Y} = \{1, \dots, C\}$ the set of C classes. Our goal is to learn a visual embedding function, $f_\theta : \mathcal{X} \to \mathbb{R}^d$, that maps semantically similar patches closer in the latent space. To provide semantic guidance, we employ a LEM, $g_\phi : \mathcal{T} \to \mathbb{R}^d$, which maps textual prompts $\mathcal{T}$ into the shared latent space. Both visual and textual embeddings are ℓ_2-normalized:

$$\mathbf{z}_x = \frac{f_\theta(x)}{\|f_\theta(x)\|_2}, \quad \mathbf{z}_t = \frac{g_\phi(t)}{\|g_\phi(t)\|_2}. \tag{1}$$

This formulation induces a Riemannian geometry [30] on the unit hypersphere $\mathbb{S}^{d-1}$, and allows the similarity between visual and textual embeddings to be measured via the cosine of the geodesic angle.

Prompt Engineering and LEM Embeddings. Each class C is assigned a single descriptive prompt t_C that narrates the distinguishing visual and semantic characteristics of hyperspectral patches, such as crop type, cultivation method, vegetation density, or aerial perspective. These prompts are designed from domain knowledge and are made to be informative, discriminative, and non-redundant across classes, so that the resulting embeddings are maximally separated in semantic space. The LEM, g_ϕ, maps each prompt to a fixed embedding. During training, the LEM is linear-probed, and the textual embeddings are kept fixed. These embeddings serve as *semantic anchors* for the CLIP-style [26] contrastive objective, and guide the visual embeddings to align with the corresponding class semantics. The prompts are designed in this manner:

> ✔ **Descriptive Prompt Template**. This image shows a large cultivated field of **{<CLS>}**, where **{<CLS>}** plants are densely grown in rows; the vivid green **{<CLS>}** vegetation is clearly visible from an aerial perspective.

Other classes use similarly structured prompts with <CLS> as the class placeholder. For the vision backbone, we employ *Masked Vision Transformer* [3, 8, 12], that we train end-to-end on hyperspectral patches using the proposed contrastive objective.

Contrastive Loss with Hard and Semi-Hard Negatives. Let $\mathbf{z}_i = \mathbf{z}_{x_i} \in \mathbb{R}^d$ denote the ℓ_2-normalized embedding of the i^{th} hyperspectral patch produced by the vision encoder, and let $\mathbf{p}_j = \mathbf{z}_{t_j} \in \mathbb{R}^d$ denote the ℓ_2-normalized textual prototype embedding for class j.

We now define a scaled cosine similarity between an image embedding and a class prototype as:

$$s_{ij} = \tau\, \mathbf{z}_i^\top \mathbf{p}_j, \quad \tau = e^{\text{logit\_scale}}, \quad i \in \mathcal{B},\, j \in \{1, \dots, C\}, \tag{2}$$

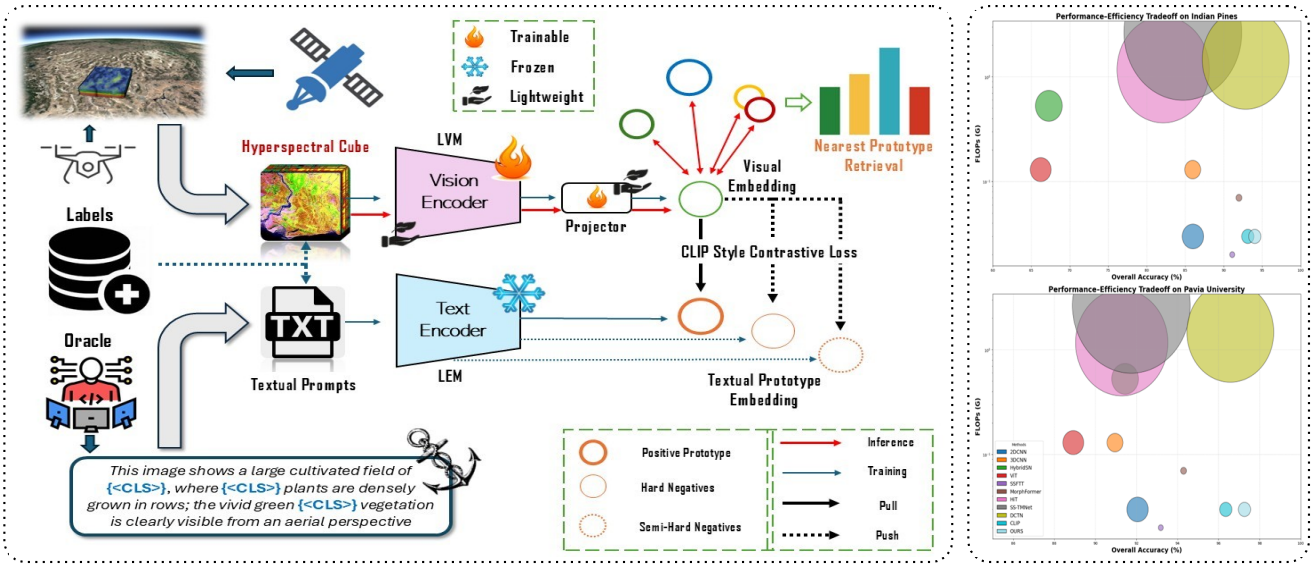
where $\tau > 0$ is a learnable temperature scaling the distribution, and $\mathcal{B}$ denotes the training batch.

Positive logit For patch i with ground-truth label y_i, the positive logit is: $s_i^+ = s_{iy_i}$, representing similarity to the correct class prototype.

Hard negatives Let k_h denote the number of top-hard negatives. For patch i, the indices of the top-k_h most confusing incorrect classes are:

$$H_i = \text{Top-}k_h\Big(\{s_{ij} \mid j \neq y_i\}\Big), \tag{3}$$

and the corresponding logits are $s_i^{\text{hard}} = \{s_{ij} \mid j \in H_i\}$.



(a) (b)

Figure 1. **Graphical snapshot of our approach**: (a) Vision–Language Model for hyperspectral scene understanding. (b) Performance–Efficiency tradeoff across different methods on IP (up) and PU (down); the size of a circle is proportional to the total parameters (MB). The x-axis denotes the overall accuracy in (%) and the y-axis represents FLOPs in G. The legend is common for both datasets.

Semi-hard negatives Let k_s denote the number of semi-hard negatives. We randomly sample k_s number of negatives from the remaining classes:

$$SH_i \subset \{1, \ldots, C\} \setminus (\{y_i\} \cup H_i), \quad |SH_i| = k_s, \tag{4}$$

with logits: $s_i^{\text{semi}} = \{s_{ij} \mid j \in SH_i\}$.

Combined logits We then concatenate the positive, hard, and semi-hard negatives to form the final logit vector:

$$\mathbf{s}_i = [\, s_i^+, \, s_i^{\text{hard}}, \, s_i^{\text{semi}} \,] \in \mathbb{R}^{1+k_h+k_s}. \tag{5}$$

The positive logit is always the first entry, followed by the hardest and then the semi-hard negatives.

Loss computation The cross-entropy loss over the combined logits is:

$$\mathcal{L}_i = -\log \frac{e^{s_i^+}}{e^{s_i^+} + \sum_{j \in H_i} e^{s_{ij}} + \sum_{j \in SH_i} e^{s_{ij}}}. \tag{6}$$

The batch-level loss is computed as:

$$\mathcal{L} = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \mathcal{L}_i. \tag{7}$$

This objective aligns each visual embedding $\mathbf{z}_i$ with its textual prototype $\mathbf{p}_j$, while hard and semi-hard negatives enforce fine-grained discrimination on the hypersphere $\mathbb{S}^{d-1}$.

✔ **Distractor-Aware Contrastive Alignment**. The vision and text embeddings are aligned via a contrastive loss on positive pairs and carefully curated hard and semi-hard negatives. Hard negatives sharpen class boundaries, while semi-hard negatives introduce variability. This optimization jointly enhances training efficiency (Table 1, 3, 7, 8) and embedding separability (Figure 2) at the same time.

Inference Procedure. During inference, a hyperspectral patch $x \in \mathcal{X}$ is passed through the trained vision backbone f_θ to obtain its ℓ_2-normalized embedding $\mathbf{z}_x \in \mathbb{R}^d$. This embedding is then compared against the fixed set of textual prototypes $\{\mathbf{p}_j\}_{j=1}^C$, derived from the class-specific prompts, using cosine similarity. The predicted class label is determined by *nearest-prototype retrieval*: $\hat{y} = \arg\max_{j \in \{1,\ldots,C\}} \mathbf{z}_x^\top \mathbf{p}_j$. This way, the classification at inference time reduces to a similarity search in the shared cross-modal embedding space, with *no additional trainable parameters required*.

3. Experimental Setups

This section outlines the experimental setup used to train on two benchmark hyperspectral datasets [1].

Datasets Used. We evaluate on two widely-used benchmark hyperspectral datasets: *Indian Pines* [1] and *Pavia University* [1]. The *Indian Pines (IP)* dataset was collected by the AVIRIS sensor over agricultural fields in Northwest-

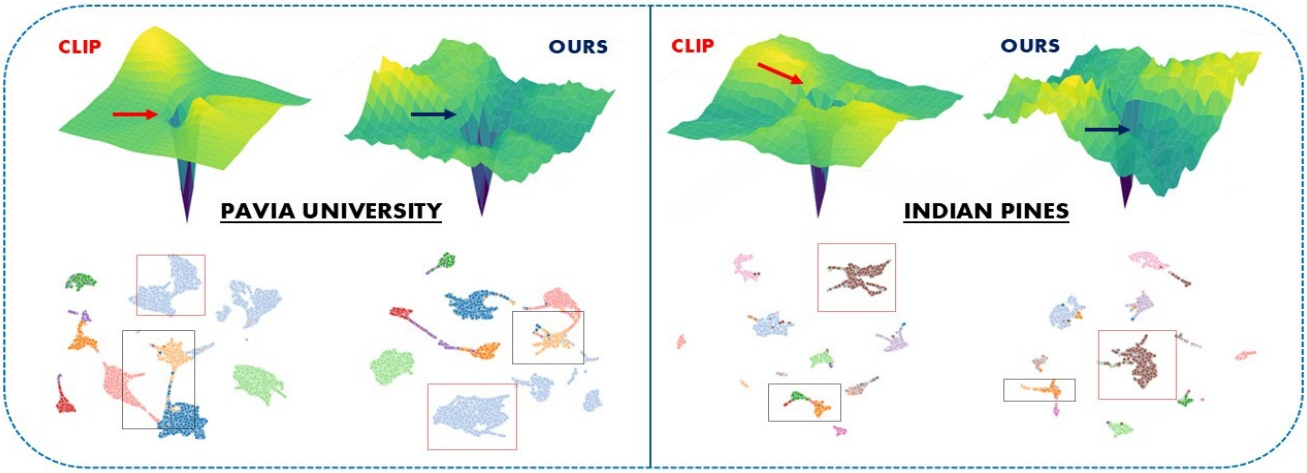


Figure 2. Loss landscape (top) and UMAP projections (bottom). From left to right: PU and IP datasets. Our model shows a more structured, generalizable landscape and neural-collapse-like embeddings, unlike the smoother but overlapping CLIP features. For instance, CLIP sometimes displays a butterfly-like structure for classes with larger samples, whereas our method exhibits a nearly circular embedding with compact intra-class relationships. Also, the inter-class separation observed is more in our case. Please zoom for better clarity.

Table 1. Comparison with other SOTA methods on various HSI datasets and their computational efficiency. **Bold**, blue, and red represent the highest/lowest, second, and third performance/efficiency. M-CLS represents multi CLS prompts.

Category	Backbone	Methods	Indian Pines		Pavia University		FLOPs (G)	Params (MB)	Latency Time (s)
			OA (%)	κ	OA (%)	κ			
Unimodal	2D CNN	2DCNN [31]	85.98	84.22	92.05	89.84	0.03	2.76	2.04
	3D CNN	3DCNN [35]	85.95	83.91	90.95	88.44	0.13	1.54	6.54
	2D + 3D CNN	HybridSN [28]	67.26	62.21	91.44	89.06	0.53	4.32	7.78
Unimodal	Vision Transformer	ViT [8]	66.21	61.65	88.92	85.81	0.13	2.61	5.31
	Token Transformer	SSFTT [32]	91.11	89.94	93.18	91.25	**0.02**	**0.15**	**1.18**
	Morph-ATT Transformer	MorphFormer [29]	91.98	90.91	94.29	92.63	0.07	0.21	9.19
	Hyperspectral Transformer	HiT [36]	82.13	79.77	91.28	88.85	1.17	51.23	11.68
	Multi-Scale Transformer	SS-TMNet [13]	84.67	82.66	91.74	89.44	2.67	83.33	31.12
	CNN + Transformer	DCTN [39]	92.85	91.87	96.57	95.49	1.48	45.32	20.69
Multimodal	Vision Language Model	M-CLS CLIP [26]	93.11	91.94	96.35	94.50	0.03	0.91	13.84
Multimodal	LEM + LVM	**OURS**	**94.03**	**93.54**	**97.26**	**96.39**	0.03	0.91	14.19
		Δ	**0.92**	**1.60**	**0.69**	**0.90**			

ern Indiana. It has a spatial size of 145×145 pixels and originally contains 220 spectral bands ranging from 0.4 to 2.5 μm. After removing water absorption bands, 200 bands are retained. The dataset is annotated into 16 land-cover classes, most of which are related to different types of crops (e.g., corn, soybeans, alfalfa), along with a few classes corresponding to natural vegetation and man-made structures. IP is particularly challenging due to its **high class imbalance**, **small sample sizes**, and the presence of spectrally similar vegetation classes. Total number of samples in the IP dataset is 10,249.

The *Pavia University (PU)* dataset was acquired by the ROSIS sensor over the University of Pavia, Italy. It has a larger spatial coverage of 610×340 pixels with 115 spectral bands, of which 103 remain after discarding noisy channels. PU contains 9 land-cover classes, including urban features (e.g., asphalt, bitumen, bricks, shadows), vegetation, and bare soils. Compared to IP, PU offers **higher spatial resolution (1.3 m per pixel)** and more spatially coherent regions, which makes it suitable for evaluating the ability of models to capture both spectral signatures and spatial context in structured urban environments. The number of samples in the PU dataset is 42,776, and it is the larger of the two.

Data Pre-Processing Pipeline Used. Given the high dimensionality and redundancy in hyperspectral data, we apply a zero padding to preserve the spatial structure at image borders and perform a Principal Component Analysis (PCA) [4] to reduce the spectral dimension to 25 principal components.

Vision Language Model Summary. The designed VLM has 335.3M parameters, with $\sim$ 335M frozen in the text encoder and only $\sim$ 240K trainable. Of these,

the vision encoder contributes 174K parameters, the projection head 65.6K, and a single logit scale parameter. Each 3×3 patch is projected into a 64-dimensional embedding and processed through 6 transformer layers with 16 self-attention heads and an MLP dimension of 64. This compact vision branch aligns with the frozen LEM (`BAAI/bge-large-en-v1.5` [2]), which provides stable 1024-D textual embeddings for descriptive prompts.

Fine-tuning Configuration. The model is trained in a vision–language alignment setup with both hard and semi-hard negatives. On Indian Pines, we train for 50 epochs with batch size 32, while on Pavia University we train for 25 epochs with batch size 128. Following the DCTN [39] protocol, 10% of labeled data is used for training and 90% for testing. Adam optimizer is used with a learning rate of 1×10^{-3}. Contrastive training uses $k_h = 4$ hard and $k_s = 4$ semi-hard negatives, with a CLIP-style loss restricted to $\{s^+, s^{\text{hard}}, s^{\text{semi}}\}$ (Sec. 2). All the experiments were simulated for four times on an NVIDIA A100 GPU.

4. Analysis of Results

In this section, we compare the proposed VLM with a set of state-of-the-art approaches and observe that it delivers performance on par with the best existing methods and yields high accuracy at a very low computational cost.

Performance Evaluation Metrics. For evaluation, we report both **Overall Accuracy (OA)** and **Cohen's Kappa coefficient (κ)**. Overall Accuracy (OA) measures the proportion of correctly classified samples across all classes, offering an indicator of classification performance. The κ coefficient, on the other hand, accounts for the agreement occurring by random chance and provides a chance-corrected measure of reliability.

Comparison with SOTA Approaches. Table 1 ellucidates the comparative performance of the proposed VLM against existing HSI classifiers. CNN-based unimodal baselines such as 2D-CNN [31], 3D-CNN [35], and HybridSN [28] deliver moderate accuracy. This is primarily due to their limited ability to capture long-range spectral–spatial dependencies. Transformer-based methods (SSFTT [32], MorphFormer [29], DCTN [39]) demonstrate clear gains and emphasize the advantage of global attention in modeling hyperspectral features. Our multimodal VLM consistently surpasses these unimodal architectures and shows the strength of cross-modal alignment. Among all the methods, DCTN [39] stands out as a hybrid architecture that combines CNNs with transformers, thus integrating both local spatial cues and long-range global dependencies. This dual design makes it the strongest vision-only backbone, capable of expressive feature learning. However, it is worth noting that optimizing unimodal vision models is often more straightforward and sometimes yields highly expressive representations, since they are not constrained by the alignment challenges posed by an additional modality. Nevertheless, even the strongest unimodal transformer backbone (DCTN) is sometimes outperformed by multimodal approaches such as CLIP [26], and this validates the importance of incorporating meaningful textual priors. CLIP [26] performs competitively versus DCTN [39] while being lighter. We observe that our method surpasses all unimodal and multimodal SOTAs on Indian Pines (+0.92 OA, +1.60 κ) and Pavia (+0.69 OA, +0.90 κ), and validates that descriptive prompts and informative negatives enable stronger cross-modal grounding.

Table 2. Ablation study of **varying loss components** on IP.

Variant	OA (%)
✗ without Hard	90.16
✗ without Semi-Hard	93.44
✔ Full Mode (Hard + Semi-Hard)	**94.03**

Parameter and Latency Analysis. Table 1 highlights the trade-off between cost and performance. CNN baselines [31, 35] are lightweight and fast but limited in accuracy. Transformer models such as SSFTT [32] offer efficiency with fewer FLOPs and parameters, while heavier ones like MorphFormer [29] and DCTN [39] demand more resources. Multimodal methods, including CLIP [26] and our method, add cross-modal alignment overhead. Yet, the proposed model achieves higher accuracy than CLIP without extra FLOPs or memory. We also notice an optimal trade-off between state-of-the-art accuracy with minimal cost and low inference time.

Ablation on Hard vs. Semi-Hard Negatives. Table 2 presents the impact of different negative sampling strategies on our contrastive alignment framework. Removing hard negatives results in a significant performance drop to 90.16% OA, and shows their importance in disentangling closely related spectral classes. Excluding semi-hard negatives leads to a higher OA of 93.44%, but still underperforms compared to the full design. Incorporating both hard and semi-hard negatives achieves the best result of 94.03% OA. From this observation in Table 2, we can infer that **hard negatives** drive the model to resolve inter-class ambiguities by pushing apart spectrally similar but semantically different categories (e.g., different crop types in Indian Pines). Meanwhile, **semi-hard negatives** act as regularizers and refine the decision boundaries. This helps in preventing the model from collapsing features of borderline or underrepresented classes. Together, their joint presence achieves both separability and robust generalization (Figure 2).

Training Data Sensitivity. Table 3 shows the sensitivity with varying supervision. Despite being one of the strongest baselines, DCTN [39] already achieves competitive overall accuracies (OA) on Indian Pines, and shows its effective

Table 3. OA (%) comparison of proposed technique with DCTN [39] on Indian Pines with varying training sample percentages.

Method	10%	20%	30%	40%	50%
DCTN [39]	92.85	95.37	95.81	96.01	96.10
OURS	**94.03**	**97.68**	**98.57**	**98.97**	**99.02**

Table 4. ablation study of **varying batch sizes** on Indian Pines.

Batch Size	4	8	16	32	64	128
OURS OA (%)	93.19	93.45	93.54	**94.03**	93.34	91.12

integration of convolutional and transformer-based modeling. However, the proposed model consistently surpasses DCTN across all training sample percentages. The improvement is seen under scarce supervision (10% training data), where our method yields a **+1.18%** OA gain (94.03% vs. 92.85%). This advantage further grows at higher data availability, and reaches near-saturation with **99.02%** OA at 50% training data. This suggests that the proposed approach not only better exploits the limited labeled samples, which is a key challenge in hyperspectral learning, but also scales more effectively as training data increases. In contrast, the performance of DCTN plateaus earlier and indicates limitations in capturing informative and more discriminative spectral–spatial features. The steady margin maintained by our model corroborates its stronger representational efficiency across varying levels of supervision.

Ablation Study on Varying Batch Size. Table 4 reports the effect of varying batch sizes on classification performance for Indian Pines. We observe a clear trade-off between representation quality and training stability. Small batch sizes (e.g., 4 or 8) provide sufficient gradient diversity but may suffer from noisier updates, and yield OA values of 93.19% and 93.45%, respectively. Increasing the batch size to 16 improves stability and achieves 93.54%, while a batch size of 32 provides the best balance and results in the highest overall accuracy of **94.03%**.

Sensitivity to Prompt Representation. The results in Table 5 depict the role of prompt design in evolving performance. When the model is guided by label-only prompts, it receives little more than a name tag and offers minimal semantic context. Short-text prompts, while an improvement, resemble terse dictionary entries that hint at meaning but fail to capture the full richness of the scene. Contrary to that, our descriptive long-text prompts act more like well-structured narratives, and try to embed both distinction and context that guide the model to anchor visual features to meaningful linguistic constructs. *It is much like how we get a deeper comprehension when immersed in a full passage rather than a single word; the model achieves stronger cross-modal alignment when exposed to richer descriptions.*

✔ **But why should a few extra words make such a difference?** Because in multimodal learning, every additional semantic pattern becomes a bridge that can tie abstract text to vision and more bridges mean stronger cross-modal alignment.

Table 5. Comparison of different **text prompt types** on the Indian Pines (IP) dataset. The results show how variations in prompt design influence classification performance.

LEM Prompt Type	**Label-only**	**Short Text**	**OURS**
OA (%)	92.90	93.07	**94.03**

Table 6. Effect of various **text-embedding backbones** on IP.

LEM Backbone	**Type**	**Family**	**OA (%)**
BAAI/bge-large-en-v1.5	English	BGE	**94.03**
BAAI/bge-M3	Multilingual	BGE	93.04
E5-Large (multilingual)	Multilingual	E5	92.72

Ablation Study on LEM Backbone Choice. Table 6 compares different text-embedding backbones used. We observe that the *English-only* BGE model (BAAI/bge-large-en-v1.5) [2] achieves the highest performance, reaching **94.03%** OA. This shows us that for hyperspectral classification tasks, where the label space is relatively small, fixed, and dominated by English terminology, a strong monolingual embedding model can provide highly discriminative representations. In contrast, the multilingual variants, although more general-purpose, exhibit slightly lower performance: BGE-M3 achieves 93.04% OA, while the E5-Large multilingual model reaches 92.72% OA. This drop may be attributed to the fact that multilingual models spread their capacity across many languages and sacrifice specialization in English, where the task-specific labels reside.

Analysis of Prompt-based vision language model with Vision-only Baseline. Comparing a descriptive prompt–based vision language model against a vision-only baseline depicts the impact of textual priors in Table 7. On both IP and PU datasets, the prompt-based approach exceeds vision-only performance (94.03% vs 91.52%, 97.26% vs 96.60%). This gap illustrates that language-derived context enhances class separability even in high-dimensional hyperspectral spaces. Moreover, the enriched semantic guidance helps resolve ambiguous spectral signatures and improves the robustness of learned embeddings across varying scenes. Figure 2 visually confirms that textual priors lead to more intra-class compactness and better inter-class margins, and also shows that textual information is not merely auxiliary but can be central in achieving high performance across datasets.

Analysis of the Loss Landscape. We analyze the training dynamics of our model with CLIP [26] in Figure 2. Loss landscape in CLIP is flat (from some directions) and smoother, while our model shows a *slightly slanted (from numerous directions)*, *rugged landscape* with a *larger minimum diameter*, and this suggests better generalization and reduced risk of getting trapped in saddle points. It is worth noting that while smooth, flat loss landscapes reduce sharp minima, but can give weaker directional optimization guidance. Excessively isotropic minima often yield Hessian spectra with compressed eigenvalue distributions, and suppress anisotropic curvature and informative descent directions. This spectral degeneracy weakens the signal-to-noise gradient signal and limits the optimizer's exploitation of the principal curvature subspaces.

Qualitative Visuals of the Embeddings. The Uniform Manifold Approximation and Projection (UMAP) [21] analysis show that our technique's embeddings exhibit a near neural-collapse-like structure [24], with tightly clustered class centroids, maximal inter-class separation, and minimal intra-class variance, contrasting with the more overlapping CLIP embeddings. Here, we attempt to analyze the UMAP projections from two complementary perspectives: (1) Latent geometry and (2) Latent separation. With respect to (1), the CLIP-induced embeddings show non-Euclidean dispersion patterns, with certain semantic manifolds exhibiting a butterfly-shaped bifurcation for PU. This effect is most conspicuous in the light-blue class (Figure 2), where the embedding unfolds into two lobes joined by a narrow topological corridor. Similarly, the light-green class also shows anisotropic curvature and is characterized by elongated eigen-directions that expose distortions in the latent geometry. Contrary to that, our approach yields manifolds that are topologically closer to isotropic Gaussian distributions, looking like circular basins in the UMAP plane. We also observe that this isotropy persists even for the deep red-labeled PU class, which otherwise exhibits higher-order deformations under CLIP. A parallel observation holds for the IP dataset also, thus ensuring robustness across datasets.

With respect to (2) Latent separation, in the CLIP embeddings of the Indian Pines (IP) dataset, we observe a noticeable inter-class connectivity involving three distinct classes, and we highlight this with a black boundary in left Sub-Figure 2 for clarity. This phenomenon reflects residual overlaps in the embedding space, where class-specific manifolds remain entangled rather than well-separated. However, our approach reduces this effect. The same region that shows entanglement under CLIP appears with reduced cross-class connectivity in our embeddings (again marked with a black dotted box). This is also observed for the Pavia dataset. This also corroborates our quantitative improvements in accuracy and robustness.

❄ **Larger Minimum Diameter?** A wider basin of convergence provides the optimizer greater flexibility and reduces sensitivity to initialization.
❄ **Slightly Rugged Landscape ($\nabla^2 L(\theta) \neq 0$)?** Local curvature, captured by the Hessian, $H(\theta) = \nabla^2 L(\theta)$, encodes fine-grained variations. Moderate eigenvalues, $\lambda_i(H) > 0$ enrich the representation space.
❄ **Slight slant from numerous directions?** If $\nabla L(\theta) \neq 0$, the loss surface is slightly tilted, i.e., there exists at least one direction of descent. When the slope is small ($\|\nabla L(\theta)\| \ll 1$), the drift is updated slowly, often biasing the trajectory toward flatter regions of the landscape. To analyze this behavior, we consider the second-order Taylor expansion: $L(\theta + \delta) \approx L(\theta) + \nabla L(\theta)^\top \delta + \frac{1}{2}\delta^\top H(\theta)\delta$, where $\theta \in \mathbb{R}^{d^P}$ denotes the parameters, δ the update step, and $H(\theta) = \nabla^2 L(\theta)$ the Hessian. Here, the first-order term $\nabla L(\theta)^\top \delta$ captures the immediate slope-driven change, while the quadratic term $\frac{1}{2}\delta^\top H(\theta)\delta$ encodes the local curvature. This clarifies how both gradient and curvature jointly determine the behavior of updates around θ.
Flatness can be quantified via the Hessian spectrum as:

$$\text{Flatness} \sim \frac{1}{n}\sum_{i=1}^{n} \lambda_i(H(\theta)),$$

where λ_i are eigenvalues and $n = \dim(\theta)$. Smaller λ_i imply broader valleys: perturbing θ along those directions changes the loss very little. Such flat regions are associated with stability of the solution and robustness to parameter noise.
Hence, flatter minima typically correspond to lower expected test loss,

$$\mathbb{E}_{x\sim\mathcal{D}}[L(\theta)] \downarrow,$$

where $\mathcal{D}$ is the data distribution. In our case, the landscape exhibits more local slants from multiple directions. This helps in converging toward the global minimum and aligns with an improved generalization performance.

Additional Benchmarking Against Pixel-Level and Training-Efficient Approaches. Table 8 highlights how recent SOTAs approach hyperspectral image classification from two distinct directions. GAF-NAU [23] improves pixel-level representation by converting 1D spectral vectors into 2D angular feature maps using Gramian Angular Field encoding, then applying a neighborhood attention U-Net to suppress irrelevant signals and strengthen class discrimination. In contrast, the Forward-Forward Algorithm (FFA) [27] targets training efficiency, replacing back-propagation

with local goodness functions to reduce computational cost and reduce vanishing gradients. Despite these advances, our patch-based approach achieves a higher accuracy on IP and PU, which shows that patch-level approaches can be a promising avenue in hyperspectral imagery.

Table 7. Comparison of **descriptive prompt–based VLM against a vision-only model** on IP and PU.

Methodology	**IP (OA %)**	**PU (OA %)**
Descriptive prompts (VLM)	**94.03**	**97.26**
Vision-only (no prompts)	91.52	96.60

Table 8. Comparison with **additional SOTAs** on IP and PU.

Methodology	*Venue*	**IP (OA%, κ)**	**PU (OA%, κ)**
GAF-NAU [23]	*CVPR'22*	81.07 / 78.31	91.12 / 88.09
FFA + BP [27]	*CVPR'24*	73.65 / 69.78	92.51 / 90.11
OURS		**94.03 / 93.54**	**97.26 / 96.39**

5. Conclusion

Our proposed vision-language model for hyperspectral image classification demonstrates that integrating descriptive textual prompts that act as anchors to visual embeddings, significantly enhances feature discriminability, class separability, and generalization, even under limited labeled data. Quantitatively, our approach achieves a substantial accuracy improvement over SOTA models on benchmark datasets with very little parameter footprint compared to leading transformer-based models. This article also highlights the effectiveness of using both hard and semi-hard negatives along with carefully engineered prompts.

Limitations and Future Works. We notice that our method has some limitations, like sensitivity to prompt design and careful selection of text descriptors. For future work, we plan to explore automated prompt optimization to further improve classification performance in cross-scene tasks for real-world application areas.

Acknowledgment

A part of this research has received support from the IEEE Geoscience and Remote Sensing Society (GRSS) under the "ProjNET" scheme.

References

[1] Hyperspectral data sets. https://lesun.weebly.com/hyperspectral-data-set.html. 3

[2] Baai/bge-large-en-v1.5. https://huggingface.co/BAAI/bge-large-en-v1.5, 2023. BAAI General Embedding (large, English-only, version 1.5). Transforms text into 1024-dimensional dense embeddings. 5, 6

[3] A. Chatterjee, S. Ghosh, and A. Ghosh. Context-aware masking and learnable diffusion-guided patch refinement in transformers via sparse supervision for hyperspectral image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCV)*, 2025. 2

[4] H. Abdi and L J. Williams. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2 (4):433–459, 2010. 4

[5] L. Alzubaidi, J. Zhang, A J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, L. Farhan, et al. Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data*, 8:1–74, 2021. 1

[6] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 35(8):1798–1828, 2013. 1

[7] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, and X. Xie. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 15(3):1–45, 2024. 2

[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. 1, 2, 4

[9] A. B. Hamida, A. Benoit, P. Lambert, and C. B. Amar. 3-d deep learning approach for remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing (TGRS)*, 56(8):4420–4434, 2018. 2

[10] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, D. Tao, et al. A survey on vision transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45(1): 87–110, 2022. 2

[11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 1

[12] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009, 2022. 2

[13] X. Huang, Y. Zhou, X. Yang, X. Zhu, and K. Wang. SS-TMNet: Spatial–spectral transformer network with multi-scale convolution for hyperspectral image classification. *Remote Sensing (RS)*, 15(5):1206, 2023. 2, 4

[14] M. J. Khan, H. S. Khan, A. Yousaf, K. Khurshid, and A. Abbas. Modern trends in hyperspectral image analysis: A review. *IEEE Access*, 6:14118–14129, 2018. 1

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1097–1105, 2012. 1

[16] D J. Lary, A H. Alavi, A H. Gandomi, and A L. Walker. Machine learning in geosciences and remote sensing. *Geoscience Frontiers*, 7(1):3–10, 2016. 1

[17] H. Lee and H. Kwon. Contextual deep CNN based hyperspectral classification. In *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 3322–3325, 2016. 2

[18] S. Li, W. Song, L. Fang, Y. Chen, P. Ghamisi, and J. A. Benediktsson. Deep learning for hyperspectral image classification: An overview. *IEEE Transactions on Geoscience and Remote Sensing (TGRS)*, 57(9):6690–6709, 2019. 1

[19] Z. Li, X. Wu, H. Du, F. Liu, H. Nghiem, and G. Shi. A survey of state of the art large vision language models: Benchmark evaluations and challenges. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1587–1606, 2025. 2

[20] B. Lu, P. D. Dao, J. Liu, Y. He, and J. Shang. Recent advances of hyperspectral imaging technology and applications in agriculture. *Remote Sensing*, 12(16):2659, 2020. 1

[21] L McInnes, J Healy, and J Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. 7

[22] H. Naveed, A. U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, and A. Mian. A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 16(5):1–72, 2025. 2

[23] S. Paheding, A. A. Reyes, A. Kasaragod, and T. Oommen. GAF-NAU: Gramian angular field encoded neighborhood attention U-Net for pixel-wise hyperspectral image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 409–417, 2022. 7, 8

[24] V. Papyan, X. Y. Han, and D. L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences (PNAS)*, 117(40):24652–24662, 2020. 7

[25] A. Plaza, J. A. Benediktsson, J. W. Boardman, J. Brazile, L. Bruzzone, G. Camps-Valls, J. Chanussot, M. Fauvel, P. Gamba, J. A. Gualtieri, M. Marconcini, J. C. Tilton, and G. Trianni. Recent advances in techniques for hyperspectral image processing. *Remote Sensing of Environment*, 113:S110–S122, 2009. 1

[26] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR, 2021. 2, 4, 5, 7

[27] A. A. Reyes and S. Paheding. Forward-forward algorithm for hyperspectral image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPR)*, 2025. 7, 8

[28] S. K. Roy, G. Krishna, S. R. Dubey, and B. B. Chaudhuri. HybridSN: Exploring 3-d–2-d CNN feature hierarchy for hyperspectral image classification. *IEEE Geoscience and Remote Sensing Letters (GRSL)*, 17(2):277–281, 2019. 2, 4, 5

[29] S. K. Roy, A. Deria, C. Shah, J. M. Haut, Q. Du, and A. Plaza. Spectral–spatial morphological attention transformer for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing (TGRS)*, 61:1–15, Art. no. 5503615, 2023. 2, 4, 5

[30] H. Shao, A. Kumar, and P. T. Fletcher. The Riemannian geometry of deep generative models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR)*, pages 315–323, 2018. 2

[31] V. Sharma, A. Diba, T. Tuytelaars, and L. Van Gool. Hyperspectral CNN for image classification & band selection, with application to face recognition. Technical Report KUL/ESAT/PSI/1604, KU Leuven, ESAT, Leuven, Belgium, 2016. 2, 4, 5

[32] L. Sun, G. Zhao, Y. Zheng, and Z. Wu. Spectral–spatial feature tokenization transformer for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing (TGRS)*, 60:1–14, 2022. 2, 4, 5

[33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5998–6008, 2017. 1

[34] J. Xiang, T. Tao, Y. Gu, T. Shu, Z. Wang, Z. Yang, and Z. Hu. Language models meet world models: Embodied experiences enhance language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 75392–75412, 2023. 2

[35] X. Yang, Y. Ye, X. Li, R. Y. K. Lau, X. Zhang, and X. Huang. Hyperspectral image classification with deep learning models. *IEEE Transactions on Geoscience and Remote Sensing (TGRS)*, 56(9):5408–5423, 2018. 2, 4, 5

[36] X. Yang, W. Cao, Y. Lu, and Y. Zhou. Hyperspectral image transformer classification networks. *IEEE Transactions on Geoscience and Remote Sensing (TGRS)*, 60:1–12, 2022. Art. no. 5528715. 2, 4

[37] J. Yoon. Hyperspectral imaging for clinical applications. *BioChip Journal*, 16(1):1–12, 2022. 1

[38] J. Zhang, J. Huang, S. Jin, and S. Lu. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 46(8): 5625–5644, 2024. 2

[39] Y. Zhou, X. Huang, X. Yang, J. Peng, and Y. Ban. DCTN: Dual-branch convolutional transformer network with efficient interactive self-attention for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing (TGRS)*, 62:1–16, 2024. 1, 2, 4, 5, 6