Multimodal Prior Learning with Double Constraint Alignment for Snapshot Spectral Compressive Imaging

Mingjin Zhang[1], **Longyi Li**[*1], **Fei Gao**[1], **Qiming Zhang**[2] and **Jie Guo**[*1]
[1]Xidian University
[2]University of Sydney
{mjinzhang, fgao}@xidian.edu.cn, lilongyi@stu.xidian.edu.cn, qzha2506@uni.sydney.edu.au, jguo@mail.xidian.edu.cn

Abstract

The objective of snapshot spectral compressive imaging reconstruction is to recover the 3D hyperspectral image (HSI) from a 2D measurement. Existing methods either focus on network architecture design or simply introduce image-level prior to the model. However, these methods lack guiding information for accurate reconstruction. Recognizing that textual description contain rich semantic information that can significantly enhance details, this paper introduces a novel framework, CAMM, which integrates text information into the model to improve the performance. The framework comprises two key components: Fine-grained Alignment Module (FAM) and Multimodal Fusion Mamba (MFM). Specifically, FAM is used to reduce the knowledge gap between the RGB domain obtained by the pre-trained vision-language model and the HSI domain. Through the double constraints of distribution similarity and entropy, the adaptive alignment of different complexity features is realized, which makes the encoded features more accurate. MFM aims to identify the guiding effect of RGB features and text features on HSI in space and channel dimensions. Instead of fusing features directly, it integrates prior at image-level and text-level prior into Mamba's state-space equation, so that each scanning step can be accurately guided. This kind of positive feedback adjustment ensures the authenticity of the guiding information. To our knowledge, this is the first text-guided model for compressive spectral imaging. Extensive experimental results the public datasets demonstrate the superior performance of CAMM, validating the effectiveness of our proposed method.

1 Introduction

Snapshot spectral compressive imaging offers significant advantages over traditional spectral imaging techniques, including cost-effectiveness, fast data acquisition, and optimized resource utilization. These benefits have driven its increasing adoption in various applications, such as remote sensing and medical imaging. Based on compressive sensing theory, the Coded Aperture Snapshot Spectral Imaging (CASSI) system [Wagadarikar *et al.*, 2008] was developed to efficiently capture hyperspectral images (HSIs) in a single shot. This is achieved by modulating the HSI signal across multiple spectral bands and integrating the modulated information into a two-dimensional compressed measurement. The Single-Dispersion CASSI (SD-CASSI) system has emerged as a research focus due to its simple architecture, with most existing algorithms tailored for this system [Cai *et al.*, 2022b; Dong *et al.*, 2023; Zhang *et al.*, 2024c; Wu *et al.*, 2025; Zhang *et al.*, 2024b]. However, it is constrained by energy limitations in sensor devices, which often compromise spatial resolution to achieve higher spectral resolution. Furthermore, the scarcity of prior information in blind imaging processes further limits the advancement of unimodal imaging algorithms. Consequently, fusion of multimodal data has become a critical strategy to enhance imaging accuracy.

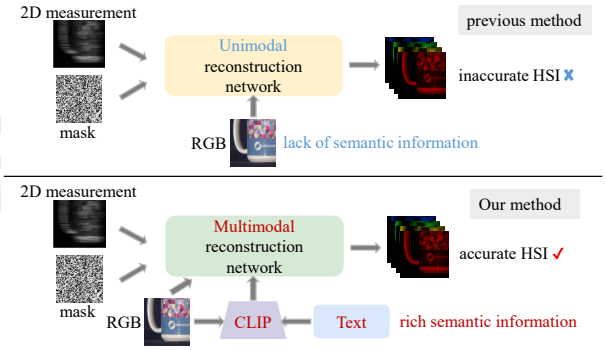

Figure 1: Comparison of our method with previous methods.

The Dual-Camera CASSI (DC-CASSI) system represents a hardware-centric solution, incorporating an additional camera to capture the scene simultaneously [Wang *et al.*, 2016; Wang *et al.*, 2015; He *et al.*, 2021]. This configuration splits the light into two paths using a beam splitter, where an uncoded grayscale or panchromatic camera is used to address ill-posed reconstruction challenges. A lot of research is starting to emerge on it. [Chen *et al.*, 2024] leverages RGB im-

*Corresponding Authors

age to enhance spatial details designs a generative model to improve spectral quality. [Cai *et al.*, 2024a] adopts spectral–spatial MLP with a network using CASSI measurements and RGB as inputs for efficiency. [Wang *et al.*, 2024a] exploits intra-inter similarity between prior image and HSI via Transformer. Despite the promising performance of the methods in image-to-image restoration, they often fail to enhance key components due to a lack of guiding information. Compared to image-level priors, text descriptions of images provide richer semantic information with not only the overall style but also the local features of main objects, which can be used as supplementary information. However, text integration into the reconstruction network faces two problems: *I. Vision-language models are trained on RGB images, but there is a gap between the hyperspectral data we use and those data. Thus, how to improve alignment between hyperspectral features and text information? II. Due to the dimensionality difference of different modalities, how to use the obtained image-level prior and text-level prior to accurately guide HSI reconstruction?*

Considering this, we propose a new multimodal learning framework, CAMM, to achieve high-quality reconstruction in this paper. It consists of two modules: Fine-grained Alignment Module (FAM) and Multimodal Fusion Mamba (MFM). Specifically, FAM reduces the bias between the pretrained Vision-Language Models (VLMs) which used to encode image and text information and the knowledge required for our tasks. Due to the different properties of features in different regions, it considers the joint constraints of distribution similarity and information entropy, and balances the optimization of simple and complex regions through adaptive strategies, making the alignment more comprehensive. This operation increases the reliability of the prior. Building on this, we propose MFM to promote the guidance of image-level prior and text-level prior to HSI. Instead of fusing the features of different modalities at one time, it integrates the features which act as the prior into the state-space equation, so that each scanning iteration of Mamba can obtain accurate guidance. This kind of positive feedback adjustment enables each position in space to obtain more accurate semantic information. In addition, we extend the public datasets, including CAVE dataset and KAIST dataset, with textual data generated by GPT-4V, with manual verification. Comprehensive experimental assessments substantiate the efficacy of CAMM, showcasing its ability to attain state-of-the-art performance in HSI reconstruction tasks.

In summary, our contributions are as follows:

- We introduce a novel framework, CAMM, which introduces text information into spectral snapshot reconstruction for the first time. Its effectiveness has been thoroughly validated through comprehensive ablation studies and comparisons with existing methods.
- A fine-grained alignment module, FAM, is designed to promote the alignment of regions with different feature complexity by the dual measurement of distribution overlap degree and information entropy.
- We design a component, MFM, in which the positive feedback adjustment mechanism improves Mamba scanning, so that image-level prior and text-level prior can provide reliable information to HSI at the global level.

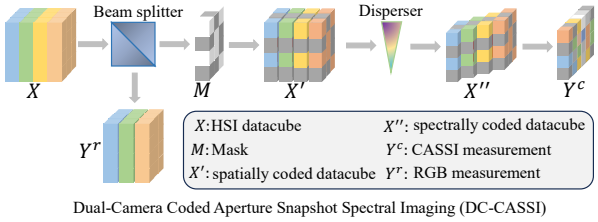

Figure 2: The formation process of the DC-CASSI.

2 Related Works

2.1 Model of DC-CASSI System

The theoretical system diagram of DC-CASSI is shown in Figure 2. A 3D HSI cube can be denoted by $X \in \mathbb{R}^{n_x \times n_y \times n_\lambda}$, where n_x, n_y, n_λ represent height, width, and number of wavelengths. First, the scene goes through a beam-splitter. If the energy of the light is evenly spilt, the RGB measurement can be written as:

$$Y_R = \frac{1}{2}\int_\Lambda \omega_{R,G,B}(\lambda) X(x,y,\lambda) d\lambda + N \tag{1}$$

where $x \in n_x$, $y \in n_y$, $\lambda \in n_\lambda$, $X(x,y,\lambda)$ is the 3D HSI cube. Λ is the spectral response range of the RGB detector and $\omega(\lambda)$ is the spectral response function, N is the noise. In CASSI system, the scene is first modulated by the mask M:

$$X'(x,y,:) = \frac{1}{2} M(x,y) \odot X(x,y,:) \tag{2}$$

where $X' \in \mathbb{R}^{n_x \times n_y \times n_\lambda}$ denotes the modulated data. Following the horizontal shift by a distance d in accordance with the dispersion process, the signal can be formulated as:

$$X''(x,y,:) = X'(x, y + d_\lambda, :) \tag{3}$$

Finally, the captured 2D compressed measurement $Y \in \mathbb{R}^{n_x \times (n_y + d(n_\lambda - 1))}$ can be obtained by:

$$Y_C = \frac{1}{2}\sum_{\lambda=1}^{n_\lambda} M(x, y + d_\lambda) \odot X(x, y + d_\lambda, :) + G \tag{4}$$

where $G \in \mathbb{R}^{n_x \times (n_y + d(n_\lambda - 1))}$ is the measurement noise. After combining Eq. 1 and Eq. 4and rewriting them as linear transformations, the imaging model can be obtained:

$$\begin{cases} Y_r = \Phi_r X + N_r \\ Y_c = \Phi_c X + N_c \end{cases} \tag{5}$$

where Φ_r is the sensing matrix of the RGB detector. Φ_c is the sensing matrix of the CASSI detector, which is usually considered as the shifted mask.

2.2 HSI Reconstruction

Various strategies have been developed to address the challenges of HSI reconstruction from DC-CASSI system. In [Wang *et al.*, 2015], the grayscale camera measurement is

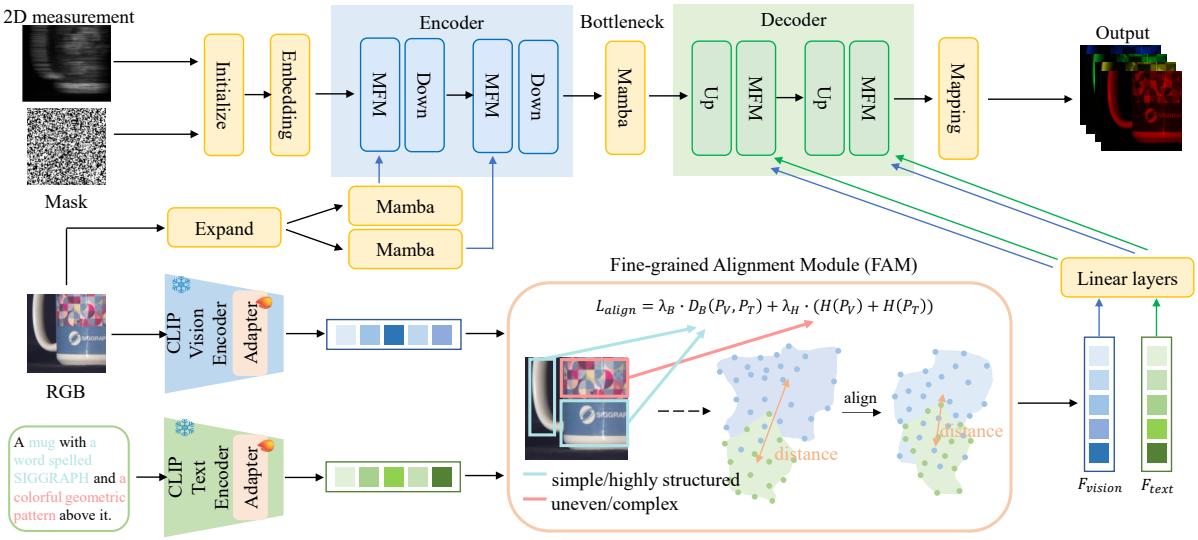

Figure 3: Overview of the proposed CAMM with FAM and MFM. The MFMs in the encoder only accept the guidance of RGB image features, while those in the decoder accept the guidance of RGB image features and text features improved by FAM.

appended to the CASSI measurement. Then the external similarity between the HSI and the panchromatic image is explored [Wang *et al.*, 2016]. A novel regularization term [Chen *et al.*, 2023] has been proposed to incorporate the semantic information from the RGB/gray measurements. [Li *et al.*, 2024a] facilitates the effective enhancement of imaging quality through the mechanism of attention, with a focus on the spatial and spectral domains. [Cai *et al.*, 2024b] embeds a dynamic mask module in front of the cross-attention-based dual-stream network to improve the quality of the reconstruction. However, there are multiple degradations that affect structural information and fine-grained details during imaging. Due to the lack of sufficient information from other modalities, the reconstruction results can easily be affected.

2.3 State Space Model

State Space Models (SSMs) have garnered significant interest due to their capability to model long-range dependencies while maintaining linear computational complexity. These models employ a continuous framework to transform a one-dimensional input signal into an output via intermediate implicit states. Mamba [Gu and Dao, 2023], an advanced SSM, is an input-dependent selection mechanism derived from structured state space models(S4) [Gu *et al.*, 2021a]. Its adaptability has led to extensive research in diverse domains such as language understanding [Gu *et al.*, 2021b] and general vision [Zhu *et al.*, 2024; Liu *et al.*, 2024b]. Recently, Mamba has also made some progress in the direction of multimodal fusion [Li *et al.*, 2024b; Zhang *et al.*, 2024a; Liu *et al.*, 2024a]. However, these studies lack the exploration of HSI reconstruction tasks. Since VLMs are usually pre-trained on RGB datasets, how to apply the learned knowledge to a different domain, HSI, needs to be explored.

3 Method

3.1 Overall Architecture

The CAMM architecture, as illustrated in Figure 3, is a U-shaped network. The process begins by inverting the measurement to recover the original shape. Following this, the mask is expanded along the channel dimension to create a 3D mask. This 3D mask is then processed through a convolution operation to generate the initial value. Both the encoder and decoder of the network contain two MFMs. The difference is that the encoder only accepts the guidance of RGB image features, while the decoder accepts the guidance of the RGB features as well as the text features. For convenience, we use vanilla Mamba as the bottleneck of the network and the model handling the RGB image. To extract vision embeddings and text embeddings, we encode RGB image and text using the pre-trained CLIP [Radford *et al.*, 2021]. For the subsequent fine-grained alignment operation of FAM, we add a kind of adapter [Wang *et al.*, 2024b] to the CLIP encoder and train it, freezing other parameters.

3.2 Fine-grained Alignment Module (FAM)

The process of FAM is shown in the lower part of Figure 3. We use CLIP, an efficient and widely used VLM, to align vision and textual features. The contrast learning of CLIP is achieved by cosine similarity between samples, which is a coarse-grained alignment. However, for our task, the network needs detailed and accurate information to guide reconstruction. In addition, the pre-trained CLIP acquires universal knowledge on a large amount of data, and when applied to a specific task, fine-tuning is required to reduce the bias between different data. Considering this, a double-constrained fine-grained alignment method is proposed.

The alignment process can be explained by the principle of energy minimization. The distribution of the two kinds of features are similar to two objects in a physical system, their sim-

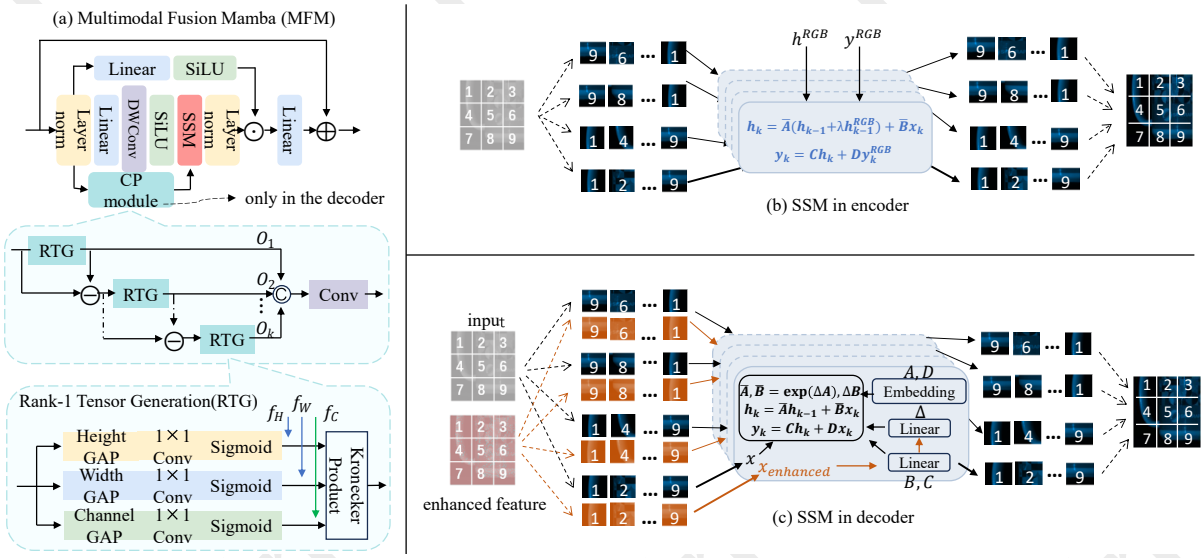

Figure 4: (a) Detailed structure of MFM. The process of SSM in encoder and decoder are shown in (b) and (c), respectively.

ilarity can be measured by Bhattacharyya distance [Kailath, 2003], while information entropy [Tsai *et al.*, 2008] measures the complexity or confusion of each distribution. Specifically, for the *vision feature probability distribution* $\mathcal{P}_V$ and the *text feature probability distribution* $\mathcal{P}_T$, the Bhattacharyya distance is defined as:

$$D_B\left(\mathcal{P}_V, \mathcal{P}_T\right) = -\ln\left(\sum_x \sqrt{\mathcal{P}_V(x)\mathcal{P}_T(x)}\right) \tag{6}$$

and the entropy of $\mathcal{P}_V$ and $\mathcal{P}_T$ is expressed as:

$$H\left(\mathcal{P}_V\right) = -\sum_x \mathcal{P}_V(x)\ln\mathcal{P}_V(x) \tag{7}$$

$$H\left(\mathcal{P}_T\right) = -\sum_x \mathcal{P}_T(x)\ln\mathcal{P}_T(x) \tag{8}$$

Under this assumption, the process of alignment between distributions can be understood as minimizing the potential energy of both by adjusting the weights. In physics, the gravitational potential energy between two objects is inversely proportional to the distance between them, and for $\mathcal{P}_V$ and $\mathcal{P}_T$, we can think of their similarity as inversely proportional to the gravitational potential energy:

$$U_B = -\frac{1}{BC\left(P_V, P_T\right)} \tag{9}$$

where $BC\left(P_V, P_T\right) = \sum_x \sqrt{\mathcal{P}_V(x)\mathcal{P}_T(x)}$ is the Bhattacharyya coefficient. When it is smaller, the potential energy is larger, indicating a large difference, so we need to pay more attention to the weight of the Bhattacharyya distance.

Using the idea of minimizing gravitational potential energy, we introduce a response parameter, k_B, and adjust the weight of this term through the exponential decay function:

$$\lambda_B = \frac{1}{1+\exp\left(-k_B\cdot\left(1-BC\left(P_v, P_T\right)\right)\right)} \tag{10}$$

a small $BC\left(P_V, P_T\right)$ means that the distribution similarity is low, and an increase in λ_B means that the Bhattacharyya coefficient contributes more to the alignment process.

In physics, free energy reflects the order or disorder of a system, and information entropy reflects the complexity of a system. In order to adaptively adjust the influence of information entropy on alignment, we define the following free energy function:

$$F_H = H\left(P_V\right) + H\left(P_T\right) \tag{11}$$

When the entropy difference $\left|H\left(P_V\right) - H\left(P_T\right)\right|$ is larger, it means that the information complexity is higher and the disorder of the system increases. To minimize energy, it is necessary to increase the focus on areas with complex information. Therefore, we combine the entropy difference and the free energy and introduce an adaptive adjustment via the exponential decay function:

$$\lambda_H = 1 - \frac{\left|H\left(P_V\right) - H\left(P_T\right)\right|}{\max\left(H\left(P_V\right), H\left(P_T\right)\right)} \cdot \frac{1}{1+\exp\left(-k_H\cdot\left|H\left(P_V\right) - H\left(P_T\right)\right|\right)} \tag{12}$$

where k_H is a hyperparameter that controls the decay rate. When $\left|H\left(\mathcal{P}_V\right) - H\left(\mathcal{P}_T\right)\right|$ is large, the weight of information entropy λ_H increases.

The Bhattacharyya distance and information entropy during alignment affect the similarity and complexity of the two distributions, and ultimately we want to minimize the total energy of the system. Therefore, we combine the weights of the two terms to construct the following loss function:

$$\mathcal{L}_{\text{align}} = \lambda_B D_B\left(\mathcal{P}_V, \mathcal{P}_T\right) + \lambda_H\left(H\left(\mathcal{P}_V\right) + H\left(\mathcal{P}_T\right)\right) \tag{13}$$

The two combine to build a more complex and comprehensive alignment constraint, allowing the model to not only look for the minimization of overlapping regions, but also to make fine adjustments to the features of complex distributions. This

combination can improve the mathematical expression of the model's feature alignment, so that it has a stronger ability to capture small differences and features in the feature space.

3.3 Multimodal Fusion Mamba (MFM)

The structure of MFM is shown in Figure 4. Because of Mamba's powerful global modeling capabilities and low complexity, we use it in the backbone network.

In the encoder, we first input RGB image into Mamba to extract features. Since RGB image retains better spatial features, we use it to supplement spatial information for HSI. To facilitate later feature fusion, we map its channel to the same dimension as the HSI. Next, we input the HSI into another Mamba to extract features. Instead of just adding the extracted RGB features to achieve enhancement, we also consider the guided process of the hidden state. For each status update during the scanning, we add the value of the hidden state obtained from the previous processing of RGB. In this way, we refine the feature fusion by step-by-step instruction, making Mamba's output more accurate each time. For this Mamba, the state space equation becomes:

$$\begin{aligned} h_k &= \overline{\mathbf{A}}(h_{k-1} + \lambda h_{k-1}^{RGB}) + \overline{\mathbf{B}} x_k \\ y_k &= \mathbf{C} h_k + \mathbf{D} y_k^{RGB} \end{aligned} \tag{14}$$

where h^{RGB} and y^{RGB} are the hidden state and output of the Mamba that deals with RGB image, respectively. λ is the hyperparameter that adjusts the degree of guidance.

In the decoder, we supplement spatial and channel information for HSI with visual and text features encoded by CLIP. Because their dimensions (1D) do not match those of HSIs (3D), it is a problem how to fuse information from different modalities. As a common tensor decomposition method, CP decomposition mines implicit relationships in data according to which we can represent the original tensor in terms of low-dimensional tensors. For an N-order tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots I_N}$, its CP decomposition can be written as:

$$\mathcal{X} = \sum_{i=1}^{k} \omega_i v_{i1} \otimes v_{i2} \otimes \cdots \otimes v_{iN} \tag{15}$$

where $v_{in} \in \mathbb{R}^{I_n}$ is a rank-1 Kronecker basis vector, ω_i is the scalar weight, and k is a predefined number called CP rank.

For HSI, the data are 3D tensors. We can implement CP decomposition and rank-1 tensor generation according to [Zhang *et al.*, 2021]. The three basis vectors $v_{i1} \in \mathbb{R}^H$, $v_{i2} \in \mathbb{R}^W$ and $v_{i3} \in \mathbb{R}^C$ contain HSI information in terms of height, width, and channel dimensions respectively. At the same time, they have the conditions to align with the dimensions of the vision and text features encoded by CLIP. Specifically, we use the linear layer to extract height and width information from vision features and channel information from text features. The process is:

$$f_H = \Psi_v^H\left(F_{\text{vision}}\right), f_W = \Psi_v^W\left(F_{\text{vision}}\right), f_C = \Psi_t^C\left(F_{\text{text}}\right) \tag{16}$$

where Ψ_v^H, Ψ_v^W, and Ψ_t^C are linear layers that extract information in three directions.

To integrate information from them and corresponding basis vectors, we utilize multi-head cross attention (MHCA). Since we are looking to match the relevant features from the extracted information, the basic vector is selected as the query. Take the height dimension for example:

$$\begin{aligned} &v_{i1} = \mathrm{LN}(\mathrm{GELU}(\mathrm{Conv}(\mathrm{v_{i1}}))), f_H = \mathrm{LN}(f_H) \\ &v_{i1} \rightarrow \mathrm{Q}, f_H \rightarrow \mathrm{K}, \mathrm{V}, v_{i1}' = \mathrm{MHCA}(\mathrm{Q}, \mathrm{K}, \mathrm{V}) \end{aligned} \tag{17}$$

where v_{i1}' is used to obtain rank-1 tensor $\mathcal{O}_i$ through Kronecker product.

Once the rank-1 component tensor are available, we aggregate them into a low-rank tensor. To learn the weights ω_i, a 3×3 convolutional layer is used to fuse different tensors:

$$\mathcal{Y}_{\text{lr}} = \mathrm{Conv}\left(\mathrm{Concat}\left(\mathcal{O}_1, \mathcal{O}_2, \ldots, \mathcal{O}_k\right)\right) \tag{18}$$

The aggregated tensor, which encapsulates comprehensive contextual details, serves as a 3D attention map that captures global correlations across various dimensions and frequency ranges. Consequently, we enhance the image features by applying the Hadamard product between this aggregated tensor and the original input features.

To enable Mamba to better capture the relationship between features in different locations during scanning, instead of using input to generate parameters $\mathbf{B}$, $\mathbf{C}$ and Δ, we use the enhanced image features to accomplish the task. In this way, the hidden state and the output obtained by each scanning step will be more accurate, and the network can obtain more accurate global information while ensuring that the input information is not lost. This positive feedback makes Mamba's output always maintain good results.

3.4 Loss Function

We adopt Charbonnier Loss [Lai *et al.*, 2018] between the output $\hat{x}$ and the clean HSI x to optimize the reconstruction:

$$\mathcal{L}_{\text{rec}} = \sqrt{\|\hat{x} - x\|^2 + \epsilon^2} \tag{19}$$

where $\epsilon = 10^{-3}$ is a constant.

For the training of the overall framework, we also take alignment loss into account. Therefore, the overall loss function can be expressed as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rec}} + \xi \mathcal{L}_{\text{align}} \tag{20}$$

where the penalty coefficient ξ is set to 0.2 by default.

4 Experiments

4.1 Experimental Settings

CAVE Dataset and KAIST Dataset. The KAIST dataset [Choi *et al.*, 2017] consists of 30 HSIs with a space size of 2704×3376. The CAVE data [Yasuma *et al.*, 2010] is of spatial size 512×512. In our experiments, adhering to the methodology outlined in [Meng *et al.*, 2020], we choose 10 scenes from the KAIST dataset to serve as the testing dataset. Each scene is cropped to a data size of $256 \times 256 \times 28$. Additionally, we select 10 scenes from the CAVE dataset, with each scene having a size of $512 \times 512 \times 28$.Following previous works [Meng *et al.*, 2020; Huang *et al.*, 2021; Cai *et al.*, 2022a], we select 28 wavelengths ranging from 450 nm to 650 nm for acquiring HSI data through spectral interpolation. To simulate spectral dispersion, a two-pixel shift is applied between neighboring spectral channels. Furthermore, we extend two datasets by incorporating corresponding textual data, offering rich semantic insights.

Algorithms	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Avg
TV-DC	35.81	33.22	31.07	40.11	33.32	34.62	31.09	32.31	29.36	33.84	33.47
	0.947	0.885	0.879	0.947	0.944	0.943	0.885	0.916	0.862	0.953	0.910
PFusion	40.09	38.84	38.70	46.65	32.07	37.12	39.74	36.75	34.52	35.53	38.00
	0.979	0.968	0.966	0.936	0.980	0.980	0.964	0.965	0.931	0.979	0.965
DAUHST	37.25	39.02	41.05	46.15	35.80	37.08	37.57	35.10	40.02	34.59	38.36
	0.958	0.967	0.971	0.983	0.969	0.970	0.963	0.966	0.970	0.956	0.967
RDLUF-MixS2	37.94	40.95	43.25	47.83	37.11	37.47	38.58	35.50	41.83	35.23	39.57
	0.966	0.977	0.979	0.990	0.976	0.975	0.969	0.970	0.978	0.962	0.974
PIDS	42.09	40.08	41.50	48.55	40.05	39.00	36.63	37.02	38.82	38.64	40.24
	0.983	0.949	0.968	0.989	0.982	0.974	0.940	0.948	0.953	0.980	0.967
MiJUN-9stg	39.26	41.78	44.31	48.53	39.30	38.22	41.00	36.72	43.84	35.56	40.86
	0.973	0.983	0.983	0.994	0.985	0.979	0.983	0.978	0.985	0.967	0.982
DAUHST-DC-9stg	41.59	45.19	43.47	48.92	40.27	41.17	40.73	40.11	43.50	41.33	42.62
	0.985	0.991	0.984	0.993	0.988	0.988	0.979	0.986	0.988	0.990	0.987
In2SET-9stg	42.56	46.42	44.55	50.63	42.01	42.49	41.59	40.53	43.83	42.33	43.69
	0.989	0.994	0.986	0.996	0.992	0.991	0.983	0.989	0.990	0.994	0.990
CasFormer	39.75	49.52	47.13	**56.29**	41.19	37.52	**47.67**	37.01	41.70	46.09	44.39
	0.993	0.987	0.989	0.979	0.992	0.970	0.969	0.939	0.983	0.998	0.980
PiE	45.54	45.20	44.61	49.81	43.73	43.32	45.17	40.65	44.26	42.08	44.44
	0.993	0.989	0.985	0.993	0.991	0.992	0.989	0.986	0.985	0.988	0.989
SSMLP	46.92	50.75	46.79	52.22	45.96	47.81	45.25	45.89	46.26	47.83	47.57
	0.996	0.998	0.990	0.998	0.997	0.998	0.991	0.997	0.993	0.998	0.996
Ours	**48.85**	53.61	**47.88**	55.13	46.99	48.57	46.35	**47.74**	48.08	**49.27**	**49.25**
	0.998	0.998	**0.993**	0.997	0.997	0.998	**0.993**	**0.998**	**0.996**	**0.999**	**0.997**

Table 1: The PSNR (upper) in dB and SSIM (lower) results on 10 scenes (S1∼S10) in KAIST. The best results are in bold.

Component	PSNR	SSIM
w/o FAM & MFM	46.77	0.991
w/o MFM	47.81	0.994
w/o FAM	48.83	0.996
CAMM	49.25	0.997

Table 2: Ablation study on individual components.

Implementation Details. In line with standard data augmentation protocols, our method incorporates random flipping and rotation strategies. The model is developed using PyTorch on a single Nvidia GeForce A800 GPU, employing the Adam optimizer with hyperparameters β_1 set to 0.9 and β_2 set to 0.999. The training consists of 300 epochs, applying a cosine annealing scheduler with linear warm-up. We configure the learning rate to 4×10^{-4} and the batch size to 4. PSNR and SSIM [Wang *et al.*, 2004] are used as our metrics. Following [Zhang *et al.*, 2021], we set the CP rank as 4.

4.2 Quantitative Results

We perform a comparative analysis of CAMM against various other techniques, including CASSI-based methods (DAUHST [Cai *et al.*, 2022b], RDLUF-MixS2 [Dong *et al.*, 2023], MiJUN [Qin *et al.*, 2025]) and DC-CASSI-based algorithms (TV-DC [Wang *et al.*, 2015], PFusion [He *et al.*, 2021], PIDS [Chen *et al.*, 2023], DAUHST-DC [Wang *et al.*, 2024a], In2SET [Wang *et al.*, 2024a], CasFormer [Li *et al.*, 2024a], PiE [Chen *et al.*, 2024], SSMLP [Cai *et al.*, 2024a]). The comparative results across 10 simulation scenes in KAIST are presented in Table 1. It can be seen that the method of adding the prior image is much better than the traditional CASSI-based method, and our method is superior to other methods with RGB images as the prior images. In addition, most of the current advanced methods use deep unfolding networks with multiple stages of iteration. By contrast, our network achieve the best performance using only end-to-end network, which proves the effectiveness of adding text messages.

4.3 Qualitative Results

As shown in Figure 5, the proposed CAMM exhibits exceptional performance, generating smoother textures and more precise edge details while preserving the spatial consistency of uniform regions. It can be seen from the spectral density curves that our method achieves the highest correlation coefficient, demonstrating superior spectral accuracy. These results highlight the efficacy of the proposed FAM and MFM.

4.4 Ablation Study

We conduct a break-down ablation experiment to explore the impact of each component, which can be seen in Table 2. After FAM and MFM are removed, the PSNR decreased by 0.42dB and 1.44dB respectively, confirming that the two modules proposed by us have a good effect on reconstruction.

Impact of the FAM. The results of models using different alignment methods on KAIST are shown in Table 3, and the errors of the results on CAVE are shown in Figure 6. CLIP represents its own contrast loss based on cosine similarity, and BC distance represents the Bhattacharyya distance. It can be seen that the performance of our FAM is higher than the other two methods and has a better recovery effect in complex

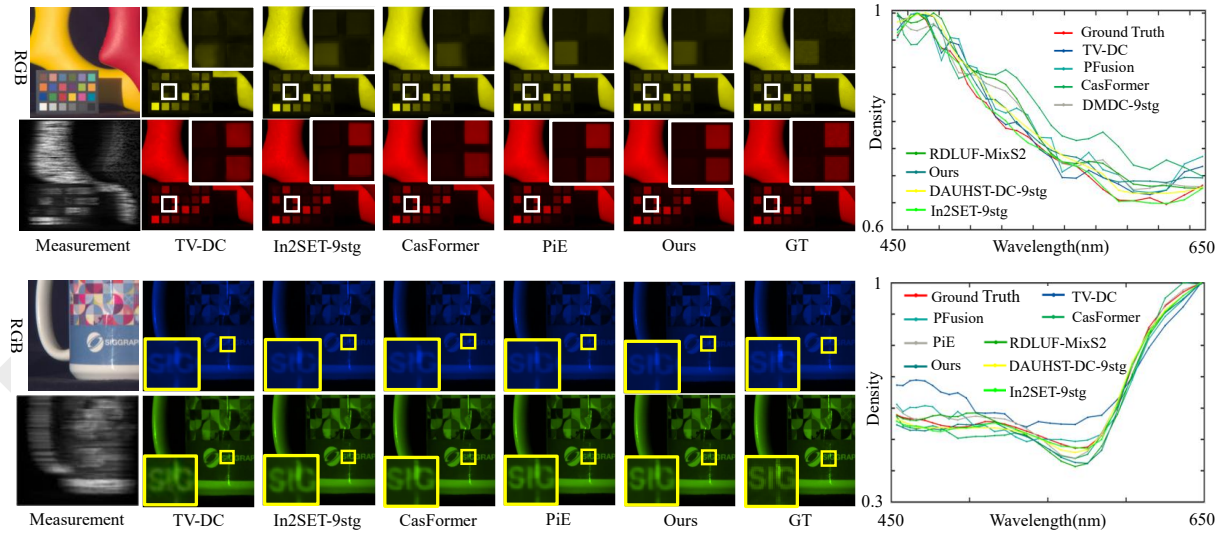

Figure 5: Visual comparsions of reconstructed HSIs in scene 3 and scene 5 with 2 spectral channels. The region within the white/yellow box is chosen for the analysis of the reconstructed spectra.

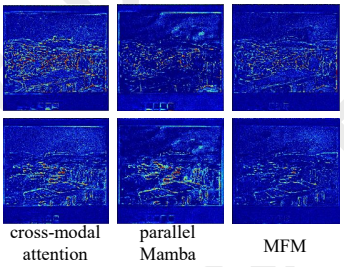

Figure 6: Residual maps between the reconstruction results and GT using different methods on the CAVE-Watercolors.

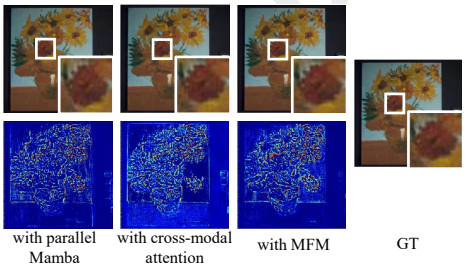

Figure 7: Visualization of HSI results using different methods on the oil-painting. The first line is the RGB image obtained with the reconstruction result, and the second line is the error map between the result and the real value on a channel.

Method	PSNR	SSIM
CLIP	48.38	0.996
CLIP + BC distance	48.87	0.996
FAM	49.25	0.997

Table 3: Comparison of different alignment methods.

Method	PSNR	SSIM
cross-modal attention	47.94	0.995
parallel Mamba	48.36	0.996
MFM	49.25	0.997

Table 4: Comparison of different multimodal fusion methods.

areas. This highlights the effectiveness of our module in implementing adaptive alignment strategies in different regions.

Impact of the MFM. The results of models using different multimodal fusion methods on KAIST are shown in Table 4, and the reconstructed results on CAVE are shown in Figure 7. Cross-modal attention represents using cross-attention for RGB-HSI and text-HSI. Parallel Mamba means letting the features of different modalities pass through Mamba separately before performing similar fusion operations. It can be seen that the performance of our MFM is higher than the other two methods and has small global error. This is due to the fact that MFM effectively blends the features of different modalities, supplementing the missing information for HSI.

5 Conclusion

In this paper, we introduce CAMM for snapshot spectral compressive imaging. Considering that image-level prior lacks rich semantic information to guide restoration, we introduce text description into the reconstruction network for the first time. FAM is proposed to further align vision-language knowledge with HSI features, adaptively processing regions of varying complexity. MFM further guides recovery using RGB and text features. By integrating image-level and text-level priors into Mamba's state update equation and output equation, we enhance features for detailed HSI reconstruction. Extensive experiments on public datasets validate our method's effectiveness.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 92470108, and Grant 62272363; in part by the Joint Laboratory for Innovation in Satellite-Borne Computers and Electronics Technology Open Fund 2023 under Grant 2024KFKT001-1, and in part by the Aeronautical Science Foundation of China under Grant 2024Z071081001.

References

[Cai *et al.*, 2022a] Yuanhao Cai, Jing Lin, Xiaowan Hu, Haoqian Wang, Xin Yuan, Yulun Zhang, Radu Timofte, and Luc Van Gool. Mask-guided spectral-wise transformer for efficient hyperspectral image reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17502–17511, 2022.

[Cai *et al.*, 2022b] Yuanhao Cai, Jing Lin, Haoqian Wang, Xin Yuan, Henghui Ding, Yulun Zhang, Radu Timofte, and Luc V Gool. Degradation-aware unfolding half-shuffle transformer for spectral compressive imaging. *Advances in Neural Information Processing Systems*, 35:37749–37761, 2022.

[Cai *et al.*, 2024a] Zeyu Cai, Ru Hong, Xun Lin, Jiming Yang, YouLiang Ni, Zhen Liu, Chengqian Jin, and Feipeng Da. A mlp architecture fusing rgb and cassi for computational spectral imaging. *Computer Vision and Image Understanding*, 249:104214, 2024.

[Cai *et al.*, 2024b] Zeyu Cai, Ziyu Zhang, Chengqian Jin, and Feipeng Da. Dmdc: a cross-attention network for dynamic mask-based dual-camera snapshot hyperspectral photography. *The Visual Computer*, pages 1–18, 2024.

[Chen *et al.*, 2023] Yurong Chen, Yaonan Wang, and Hui Zhang. Prior image guided snapshot compressive spectral imaging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):11096–11107, 2023.

[Chen *et al.*, 2024] Yurong Chen, Yaonan Wang, and Hui Zhang. Prior images guided generative autoencoder model for dual-camera compressive spectral imaging. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[Choi *et al.*, 2017] Inchang Choi, MH Kim, D Gutierrez, DS Jeon, and G Nam. High-quality hyperspectral reconstruction using a spectral prior. Technical report, 2017.

[Dong *et al.*, 2023] Yubo Dong, Dahua Gao, Tian Qiu, Yuyan Li, Minxi Yang, and Guangming Shi. Residual degradation learning unfolding framework with mixing priors across spectral and spatial for compressive spectral imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22262–22271, 2023.

[Gu and Dao, 2023] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

[Gu *et al.*, 2021a] Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021.

[Gu *et al.*, 2021b] Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021.

[He *et al.*, 2021] Wei He, Naoto Yokoya, and Xin Yuan. Fast hyperspectral image recovery of dual-camera compressive hyperspectral imaging via non-iterative subspace-based fusion. *IEEE Transactions on Image Processing*, 30:7170–7183, 2021.

[Huang *et al.*, 2021] Tao Huang, Weisheng Dong, Xin Yuan, Jinjian Wu, and Guangming Shi. Deep gaussian scale mixture prior for spectral compressive imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16216–16225, 2021.

[Kailath, 2003] Thomas Kailath. The divergence and bhattacharyya distance measures in signal selection. *IEEE transactions on communication technology*, 15(1):52–60, 2003.

[Lai *et al.*, 2018] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Fast and accurate image super-resolution with deep laplacian pyramid networks. *IEEE transactions on pattern analysis and machine intelligence*, 41(11):2599–2613, 2018.

[Li *et al.*, 2024a] Chenyu Li, Bing Zhang, Danfeng Hong, Jun Zhou, Gemine Vivone, Shutao Li, and Jocelyn Chanussot. Casformer: Cascaded transformers for fusion-aware computational hyperspectral imaging. *Information Fusion*, 108:102408, 2024.

[Li *et al.*, 2024b] Yan Li, Yifei Xing, Xiangyuan Lan, Xin Li, Haifeng Chen, and Dongmei Jiang. Alignmamba: Enhancing multimodal mamba with local and global cross-modal alignment. *arXiv preprint arXiv:2412.00833*, 2024.

[Liu *et al.*, 2024a] Chang Liu, Xin Ma, Xiaochen Yang, Yuxiang Zhang, and Yanni Dong. Como: Cross-mamba interaction and offset-guided fusion for multimodal object detection. *arXiv preprint arXiv:2412.18076*, 2024.

[Liu *et al.*, 2024b] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, and Yunfan Liu. Vmamba: Visual state space model. *arXiv preprint arXiv:2401.10166*, 2024.

[Meng *et al.*, 2020] Ziyi Meng, Jiawei Ma, and Xin Yuan. End-to-end low cost compressive spectral imaging with spatial-spectral self-attention. In *European conference on computer vision*, pages 187–204. Springer, 2020.

[Qin *et al.*, 2025] Mengjie Qin, Yuchao Feng, Zongliang Wu, Yulun Zhang, and Xin Yuan. Detail matters: Mamba-inspired joint unfolding network for snapshot spectral compressive imaging. *arXiv preprint arXiv:2501.01262*, 2025.

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

[Tsai *et al.*, 2008] Du-Yih Tsai, Yongbum Lee, and Eri Matsuyama. Information entropy measure for evaluation of image quality. *Journal of digital imaging*, 21:338–347, 2008.

[Wagadarikar *et al.*, 2008] Ashwin Wagadarikar, Renu John, Rebecca Willett, and David Brady. Single disperser design for coded aperture snapshot spectral imaging. *Applied optics*, 47(10):B44–B51, 2008.

[Wang *et al.*, 2004] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[Wang *et al.*, 2015] Lizhi Wang, Zhiwei Xiong, Dahua Gao, Guangming Shi, and Feng Wu. Dual-camera design for coded aperture snapshot spectral imaging. *Applied optics*, 54(4):848–858, 2015.

[Wang *et al.*, 2016] Lizhi Wang, Zhiwei Xiong, Guangming Shi, Feng Wu, and Wenjun Zeng. Adaptive nonlocal sparse representation for dual-camera compressive hyperspectral imaging. *IEEE transactions on pattern analysis and machine intelligence*, 39(10):2104–2111, 2016.

[Wang *et al.*, 2024a] Xin Wang, Lizhi Wang, Xiangtian Ma, Maoqing Zhang, Lin Zhu, and Hua Huang. In2set: Intra-inter similarity exploiting transformer for dual-camera compressive hyperspectral imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24881–24891, 2024.

[Wang *et al.*, 2024b] Yuhao Wang, Xuehu Liu, Tianyu Yan, Yang Liu, Aihua Zheng, Pingping Zhang, and Huchuan Lu. Mambapro: Multi-modal object re-identification with mamba aggregation and synergistic prompt. *arXiv preprint arXiv:2412.10707*, 2024.

[Wu *et al.*, 2025] Zongliang Wu, Ruiying Lu, Ying Fu, and Xin Yuan. Latent diffusion prior enhanced deep unfolding for snapshot spectral compressive imaging. In *European Conference on Computer Vision*, pages 164–181. Springer, 2025.

[Yasuma *et al.*, 2010] Fumihito Yasuma, Tomoo Mitsunaga, Daisuke Iso, and Shree K Nayar. Generalized assorted pixel camera: postcapture control of resolution, dynamic range, and spectrum. *IEEE transactions on image processing*, 19(9):2241–2253, 2010.

[Zhang *et al.*, 2021] Shipeng Zhang, Lizhi Wang, Lei Zhang, and Hua Huang. Learning tensor low-rank prior for hyperspectral image reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12006–12015, 2021.

[Zhang *et al.*, 2024a] Guanglian Zhang, Zhanxu Zhang, Jiangwei Deng, Lifeng Bian, and Chen Yang. S 2 cross-mamba: Spatial–spectral cross-mamba for multimodal remote sensing image classification. *IEEE Geoscience and Remote Sensing Letters*, 2024.

[Zhang *et al.*, 2024b] Jiancheng Zhang, Haijin Zeng, Yongyong Chen, Dengxiu Yu, and Yin-Ping Zhao. Improving spectral snapshot reconstruction with spectral-spatial rectification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25817–25826, 2024.

[Zhang *et al.*, 2024c] Mingjin Zhang, Longyi Li, Wenxuan Shi, Jie Guo, Yunsong Li, and Xinbo Gao. Vmambasci: Dynamic deep unfolding network with mamba for compressive spectral imaging. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6549–6558, 2024.

[Zhu *et al.*, 2024] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024.