TRANSITIVE SEMANTIC ALIGNMENT IN CLASS-CONDITIONED LATENT DIFFUSION FOR MARTIAN HYPERSPECTRAL MINERAL MAPPING

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

Ashish Ghosh
IIIT Bhubaneswar, India
ash@isical.ac.in

Emmett Ientilucci
Rochester Institute of Technology, USA
ejipci@rit.edu

ABSTRACT

Hyperspectral data possess rich but fragile geometric structure that is often obscured by noise, redundancy, and limited supervision. We propose *Class-Conditioned Latent Diffusion Networks (CC-LDNs)*, which cast representation learning as a reverse-time diffusion process on a latent manifold. A lightweight 3D encoder maps hyperspectral patches to latent space, where classification constraints imposed on noisy latents regularize reverse diffusion dynamics. A label- and time-conditioned denoiser recovers clean representations via *transitive semantic alignment*. Geometrically, CC-LDNs induce class-conditional basins of attraction that contract intra-class variability while preserving inter-class separation, yielding stable and well-conditioned latent structures. CC-LDN achieves OA of 96.51%, 92.37%, and 90.85% on the HC, NF, and UP Martian datasets, respectively, outperforming eight state-of-the-art methods with fewer parameters than transformer-based models, validating latent diffusion as a geometry-grounded paradigm for hyperspectral learning under limited supervision.

1 INTRODUCTION

Hyperspectral imagery (HSI) (Plaza et al., 2009; Khan et al., 2018) captures dense spectral measurements across contiguous wavelength bands, enabling fine-grained material discrimination (Healey & Slater, 1999). In planetary science, orbital instruments on **Mars**, notably **CRISM** (Murchie et al., 2007) and **OMEGA** (Bibring et al., 2006), have produced high-dimensional datasets for large-scale mineralogical mapping related to aqueous alteration and volcanism. However, automated analysis of Martian HSI remains challenging (Bioucas-Dias et al., 2013) due to extreme spectral dimensionality, sensor and atmospheric noise, and spatially varying illumination, while labeled data are scarce and costly to obtain. In contrast to terrestrial hyperspectral remote sensing, where abundant annotations enable supervised deep learning (Plaza et al., 2009; Khan et al., 2018; Chatterjee & Ghosh, 2025), Martian hyperspectral analysis is fundamentally limited by weak supervision. This setting motivates learning frameworks that can induce class-discriminative structure from noisy, high-dimensional data without direct supervision on clean representations. In particular, we seek mechanisms by which semantic constraints imposed on perturbed or intermediate representations can propagate *transitively* to clean latent embeddings through structured generative dynamics. We therefore review prior work on hyperspectral learning for planetary remote sensing below.

Traditional Deep Learning Techniques. Early Hyperspectral image classification (HSIC) methods mainly employed CNNs. Yu et al. (Yu et al., 2017) introduced 2D CNNs for spectral–spatial feature extraction, later extended to 3D CNNs by Li et al. (Li et al., 2017) for joint spectral–spatial modeling, while Zhong et al. proposed SSRN (Zhong et al., 2018) to enhance discriminative learning.

Attention-based Methods. Attention mechanisms such as squeeze-and-excitation (SE) (Hu et al., 2018) and efficient channel attention (ECA) (Wang et al., 2020) enhance HSIC by emphasizing

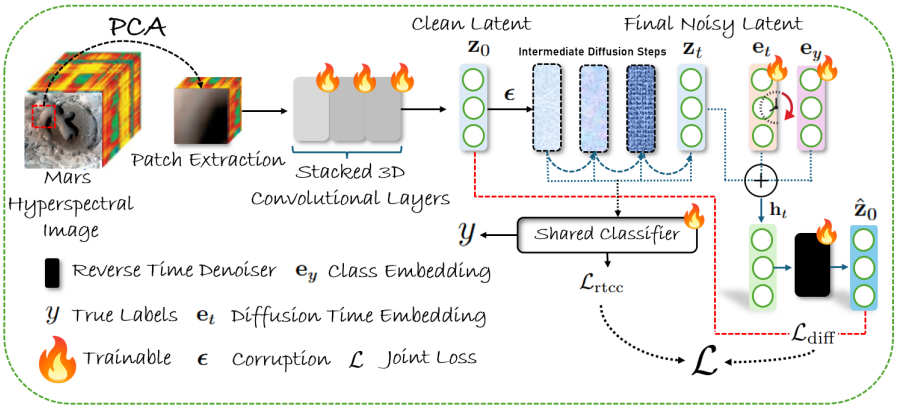

Figure 1: **Flow diagram of class-conditioned latent diffusion network *(CC-LDN)*.**

informative features (Dosovitskiy et al., 2021; Vaswani et al., 2017). Notable methods include A2S^{2}K-ResNet (Roy et al., 2021); however, CNN-based attention remains constrained by local receptive fields and limits global context.

Transformer-based Frameworks. Transformer architectures (Dosovitskiy et al., 2021; Vaswani et al., 2017; Chatterjee et al., 2025) have gained increasing attention due to their ability to capture long-range dependencies via multi-head self-attention (MHSA). Representative approaches include SpectralFormer (Hong et al., 2022), which employs group-wise spectral embedding and adaptive fusion, the group-aware hierarchical transformer (GAHT) (Mei et al., 2022) for localized spectral–spatial attention, and SPRLT-Net (Xue et al., 2022), which models fine-grained spatial relationships by treating each pixel as a token.

Hybrid Techniques. CNN–Transformer hybrids combine local feature extraction with global context modeling. Representative methods include SSFTT (Sun et al., 2022), CDSFT (Qiu et al., 2023), and MorphFormer (Roy et al., 2023). More recently, MCTGCL (Xi et al., 2025a) integrates spectral–spatial modeling, attention, morphological features, and graph contrastive learning for Martian HSI classification. In contrast to these approaches, our goal is to understand how clean latent representations can acquire class-discriminative structure without direct supervision. To this end, we impose classification constraints on noisy latent states and learn a label-conditioned reverse diffusion that maps these structured representations back to clean embeddings, enabling semantic information to propagate transitively from noisy to clean space. **Our contributions are three-fold:**

- We formulate hyperspectral representation learning as a *class-conditioned latent diffusion process*, allowing discriminative structure to emerge through reverse-time denoising rather than direct latent supervision.
- We introduce *reverse-time classification consistency*, which enforces semantic invariance across noisy latent states and induces transitive alignment between corrupted and clean representations.
- Experiments on multiple Mars hyperspectral datasets show *SOTA* mineral classification performance with fewer parameters than recent transformer-based method (Xi et al., 2025a).

The rest of the paper is organized as follows. Section 2 describes the method, Section 3 presents results, and Section 4 concludes. Additional details, methodology, and results are in Appendix A.

2 Methodology

In this section, we present *CC-LDN* (Fig. 1). We (1) encode each hyperspectral patch into a latent space using a 3D CNN, (2) perturb it via forward latent diffusion, (3) perform class- and time-conditioned reverse diffusion for reconstruction, and (4) jointly enforce classification and diffusion objectives to learn robust, class-discriminative representations for mapping. These steps are listed below in detail.

Hyperspectral Encoder. Given a hyperspectral patch $\mathbf{x} \in \mathbb{R}^{H\times W\times B}$ centered at a target pixel, with H, W, B as its height, width, and spectral dimension, an encoder (a 3D CNN) is employed to

extract joint spectral–spatial features. Three stacked 3D convolutional layers with ReLU activations transform the input into high-level representations, $\mathbf{h}$, which are aggregated into a compact latent embedding via global average pooling ($\mathrm{GAP3D}(\cdot)$) and a fully connected layer resulting into a clean latent vector $\mathbf{z}_0$, which will be further processed for diffusion mechanism:

$$\mathbf{z}_0 = \mathrm{LN}\Big(\mathbf{W}^{(f)}\,\mathrm{GAP3D}(\mathbf{h}) + \mathbf{b}^{(f)}\Big). \tag{1}$$

Here, $\mathbf{W}^{(f)}$ and $\mathbf{b}^{(f)}$ are learnable projection parameters, and $\mathrm{LN}(\cdot)$ denotes layer normalization.

Forward Latent Diffusion Process. We define a forward diffusion process (Chatterjee et al., 2025; Croitoru et al., 2023; Feng et al., 2023) in latent space to model stochastic degradation of hyperspectral representations.

Given a clean latent embedding $\mathbf{z}_0 \in \mathbb{R}^d$, Gaussian noise is injected at diffusion step t as:

$$\mathbf{z}_t = \sqrt{\bar{\alpha}_t}\,\mathbf{z}_0 + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{2}$$

where

$$\bar{\alpha}_t = \prod_{i=1}^{t}(1-\beta_i), \quad t \in \{1, 2, \ldots, T\}, \tag{3}$$

and $\{\beta_t\}$ defines a monotonically increasing noise schedule. This allows direct sampling of $\mathbf{z}_t$ at arbitrary diffusion steps.

Label-Conditioned Reverse-Time Latent Diffusion. We recover class-discriminative latent representations by learning a label-conditioned reverse-time diffusion process. Specifically, given a noisy latent $\mathbf{z}_t$ at diffusion step t and class label $y \in \{1, 2, \ldots, C\}$, a lightweight latent-space neural denoiser predicts the corresponding clean latent embedding. Formally, we define the denoising operation as:

$$\hat{\mathbf{z}}_0 = g_\phi(\mathbf{h}_t), \tag{4}$$

where $g_\phi(\cdot)$ denotes the denoiser parameterized by ϕ, and $\mathbf{h}_t$ is a conditioned latent representation discussed below.

To associate *temporal* and *semantic* information, we employ explicit embeddings for the *diffusion step* and *class label*. The diffusion time embedding $\mathbf{e}_t$ is obtained by mapping the normalized diffusion index through a multilayer perceptron as $\mathbf{e}_t = \psi_t\big(\frac{t}{T}\big)$, while the class embedding $\mathbf{e}_y$ is produced via a trainable embedding lookup, $\mathbf{e}_y = \psi_y(y)$, with $\psi(.)$ being the mapping function. These embeddings are fused additively with the noisy latent to form a conditioned representation, $\mathbf{h}_t = \mathbf{z}_t + \mathbf{e}_t + \mathbf{e}_y$, which serves as the input to the denoiser in eq. equation 4. This additive conditioning allows the denoiser to adapt its reconstruction behavior across diffusion steps while preserving class-specific semantics. The reverse-time diffusion objective is optimized by minimizing reconstruction error between the predicted clean latent and the original latent embedding across diffusion steps:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_t\big[\|\mathbf{z}_0 - \hat{\mathbf{z}}_0\|_2^2\big], \tag{5}$$

where the expectation is taken over a randomly sampled diffusion index t. By explicitly modeling class-conditioned denoising trajectories, the proposed formulation enables progressive semantic refinement of hyperspectral representations from noisy latent states, leading to improved robustness and discriminability under limited labeled data.

Reverse-Time Classification Consistency. To preserve semantic consistency along the diffusion trajectory, we employ a shared classifier, $\mathcal{C}_{shared}(\cdot)$ across latent states. The classifier is implemented as a lightweight fully connected layer with softmax activation that maps latent embeddings to class probabilities, $\mathbf{p} = \mathcal{C}_{shared}(\mathbf{z})$, and its parameters are shared for noisy latent representations during training. The reverse-time consistency is enforced by penalizing classification errors on noisy latent embeddings $\mathbf{z}_t$ sampled at arbitrary steps:

$$\mathcal{L}_{\text{rtcc}} = \mathcal{L}_{\text{ce}}(\mathcal{C}_{shared}(\mathbf{z}_t), y), \tag{6}$$

where $\mathcal{L}_{\text{ce}}(\cdot)$ denotes the cross-entropy loss and y is the ground-truth class label.

Transitive Semantic Alignment. Although explicit classification supervision is applied only to noisy latent variables, semantic structure is propagated to the denoised embedding through a transitive

alignment mechanism. A shared classifier enforces class-discriminative structure on perturbed latents $\mathbf{z}_t$, while the label-conditioned reverse diffusion model maps $\mathbf{z}_t$ to a denoised estimate $\hat{\mathbf{z}}_0$ that reconstructs the clean latent $\mathbf{z}_0$. This coupling implicitly aligns $\hat{\mathbf{z}}_0$ with the classifier-induced decision geometry, enabling semantic consistency without direct supervision on the denoised representation. We formalize the transitive semantic alignment mechanism induced by reverse-time classification consistency and label-conditioned latent diffusion in Appendix A.2.

From a geometric perspective, reverse diffusion contracts intra-class variability along diffusion trajectories while preserving inter-class separation. As a result, the clean latent space is implicitly regularized toward class-conditional manifolds with stable and well-separated decision regions, explaining the observed robustness and low variance for downstream classification in Table 4.

Overall Objective. The framework is trained end-to-end by jointly optimizing diffusion reconstruction (equation 5) and reverse-time classification consistency (equation 6):

$$\mathcal{L} = \mathcal{L}_{\text{diff}} + \lambda_{\text{rtcc}}\mathcal{L}_{\text{rtcc}}, \tag{7}$$

where λ_{rtcc} is scalar hyperparameter that balances semantic consistency. Here, λ_{rtcc} is set to 1.2.

3 RESULTS

Comparison with SOTA Methods. As summarized in Table 1, the proposed method consistently outperforms SOTA CNN/Transformer-based approaches across all three datasets. On HC, it achieves an overall accuracy (OA) of **96.51**% and a κ coefficient of **94.89**, surpassing recent models such as MCTGCL (Xi et al., 2025a) and GAHT (Mei et al., 2022). For the mineralogically complex NF data, our method attains the highest OA of **92.37**% and κ of **90.41**, and shows robustness to strong spectral variability. On UP, where class separability is more subtle, our approach achieves an average accuracy (AA) of **90.64**%, and is seen to outperform transformer-based architectures.

Computational Analysis. In addition to improved accuracy, the proposed framework exhibits favorable computational efficiency, with approximately 148K trainable parameters (Table 2 in Appendix A.3). This places the model in a **moderate** complexity regime, comparable to lightweight CNN-based approaches and substantially more compact than larger architectures such as GAHT (Mei et al., 2022) and MCTGCL (Xi et al., 2025a). Training time is also significantly lower than several transformer-based techniques.

Table 1: Comparison of our method with state-of-the-art approaches on three Martian hyperspectral datasets. Average performance over **10** runs are reported. Best results are highlighted in bold.

Methods	Venue	Holden Crater			Nili Fossae			Utopia Planitia		
		OA (%)	AA (%)	κ	OA (%)	AA (%)	κ	OA (%)	AA (%)	κ
3DCNN	*RS'17*	80.39	85.08	72.53	62.76	69.80	54.77	78.04	81.29	70.85
SSRN	*TGRS'18*	87.89	91.42	82.76	73.19	77.34	67.02	80.89	86.16	74.44
A^2S^2K	*TGRS'21*	79.97	86.91	72.11	71.20	76.49	64.72	82.06	87.27	75.97
Spectral-F	*TGRS'22*	82.33	86.91	75.14	72.21	74.92	65.79	79.47	84.50	72.56
GAHT	*TGRS'22*	87.28	89.63	81.81	74.25	76.25	68.42	82.18	85.51	76.01
SSFTT	*TGRS'22*	86.32	90.23	80.50	71.91	76.02	65.44	83.56	86.86	77.85
Morph-F	*TGRS'23*	81.47	87.14	73.73	63.11	70.31	55.41	81.16	85.25	74.63
MCTGCL	*TGRS'25*	91.99	93.30	88.39	81.20	82.68	76.57	86.93	88.68	82.19
OURS (CC-LDN)		**96.51**	**95.26**	**94.89**	**92.37**	**90.08**	**90.41**	**90.85**	**90.64**	**87.42**

4 CONCLUSION

We introduced *Class-Conditioned Latent Diffusion Networks (CC-LDNs)* for Martian hyperspectral analysis, which learn noise-robust and class-discriminative representations via reverse-time latent diffusion under limited supervision. By enforcing semantic consistency on noisy latent states, CC-LDNs enable transitive propagation of class structure to clean embeddings and induce well-conditioned, class-aligned latent geometry. Experiments on three Martian hyperspectral datasets demonstrate consistent gains over CNN- and transformer-based baselines with strong parameter efficiency, while ablations and geometric analyses reveal class-specific basins of attraction and stable reverse diffusion dynamics. The current framework employs a fixed latent dimensionality and noise schedule, which may limit adaptability across heterogeneous planetary datasets, particularly when extending beyond Mars to bodies such as Venus, where surface composition, atmospheric interference, and sensing conditions differ substantially. Future work will explore adaptive diffusion scheduling and the incorporation of mineralogical priors. **Additional analyses and visualizations are deferred to Appendix A.**

REFERENCES

Hervé Abdi and Lynne J. Williams. Principal Component Analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4):433–459, 2010.

Jean-Pierre Bibring, Yves Langevin, John F. Mustard, François Poulet, Raymond Arvidson, Aline Gendrin, and Gerhard Neukum. Global mineralogical and aqueous mars history derived from OMEGA/Mars express data. *Science*, 312(5772):400–404, 2006.

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.

Abhiroop Chatterjee and Susmita Ghosh. Learning hyperspectral images with curated text prompts for efficient multimodal alignment. *arXiv preprint arXiv:2509.22697*, 2025.

Abhiroop Chatterjee, Susmita Ghosh, and Ashish 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*, pp. 2906–2915, 2025.

Florin A. Croitoru, Vlad Hondru, Radu T. Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10850–10869, 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2021.

Bingchen T. Feng, John Smith, Michael Rubinstein, Huiwen Chang, Katherine L. Bouman, and William T. Freeman. Score-based diffusion models as principled priors for inverse imaging. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10520–10531, 2023.

Glenn Healey and Dan Slater. Models and methods for automated material identification in hyperspectral imagery acquired under unknown illumination and atmospheric conditions. *IEEE Transactions on Geoscience and Remote Sensing*, 37(6):2706–2717, 1999.

Danfeng Hong et al. SpectralFormer: Rethinking hyperspectral image classification with transformers. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 2022.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7132–7141, July 2018.

O. M. Kamps, R. D. Hewson, F. J. A. van Ruitenbeek, and F. D. Van Der Meer. Defining surface types of mars using global crism summary product maps. *Journal of Geophysical Research: Planets*, 125(8):e2019JE006337, 2020.

Muhammad Javed Khan, Hammad Shahid Khan, Adeel Yousaf, Kashif Khurshid, and Ahsan Abbas. Modern trends in hyperspectral image analysis: A review. *IEEE Access*, 6:14118–14129, 2018.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Yushi Li, Haokui Zhang, and Qiang Shen. Spectral–spatial classification of hyperspectral imagery with 3D convolutional neural network. *Remote Sensing*, 9(1):67, January 2017.

Shanshan Mei, Chao Song, Min Ma, and Feng Xu. Hyperspectral image classification using group-aware hierarchical transformer. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 2022.

Scott Murchie, Raymond Arvidson, Philippe Bedini, Kirk Beisser, Jean-Pierre Bibring, Janice Bishop, and Michael Wolff. Compact reconnaissance imaging spectrometer for mars (CRISM) on mars reconnaissance orbiter (MRO). *Journal of Geophysical Research: Planets*, 112(E5), 2007.

Antonio Plaza, Jon Atli Benediktsson, Joseph W. Boardman, Justin Brazile, Lorenzo Bruzzone, Gustau Camps-Valls, and Giuseppe Trianni. Recent advances in techniques for hyperspectral image processing. *Remote Sensing of Environment*, 113:S110–S122, 2009.

Zhen Qiu, Jie Xu, Jun Peng, and Wei Sun. Cross-channel dynamic spatial–spectral fusion transformer for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 61, 2023.

Sankar K. Roy, Subhankar Manna, Tiejun Song, and Lorenzo Bruzzone. Attention-based adaptive spectral–spatial kernel ResNet for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 59(9):7831–7843, September 2021.

Sankar K. Roy, Abhijit Deria, Chintan Shah, Juan M. Haut, Qian Du, and Antonio Plaza. Spectral–spatial morphological attention transformer for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 61:5503615, 2023.

Lei Sun, Guang Zhao, Yuhui Zheng, and Zhiwei Wu. Spectral–spatial feature tokenization transformer for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. ECA-Net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11534–11542, June 2020.

Bobo Xi, Yun Zhang, Jiaojiao Li, Tie Zheng, Xunfeng Zhao, Haitao Xu, Changbin Xue, Yunsong Li, and Jocelyn Chanussot. MCTGCL: Mixed CNN–Transformer for mars hyperspectral image classification with graph contrastive learning. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–14, 2025a.

Bobo Xi, Yun Zhang, Jiaojiao Li, et al. HyMars: Mars hyperspectral image classification benchmark datasets, 2025b. URL `https://doi.org/10.57760/sciencedb.19732`.

Zhaohui Xue, Qiang Xu, and Meng Zhang. Local transformer with spatial partition restore for hyperspectral image classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:4307–4325, 2022.

Shuang Yu, Sen Jia, and Chunyan Xu. Convolutional neural networks for hyperspectral image classification. *Neurocomputing*, 219:88–98, January 2017.

Zilong Zhong, Jun Li, Zhiwei Luo, and Mark 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, February 2018.

A Appendix

This appendix contains the supplementary materials of our manuscript and is organized as follows:

- Appendix A.1 details the experimental setups, including datasets, preprocessing procedures, and training configurations.
- Appendix A.2 presents extended methodological analysis, covering theoretical formalizations, geometric interpretations of reverse-time diffusion, and supporting propositions.
- Appendix A.3 reports additional experimental results, including scaling behavior, ablation studies, loss landscape analysis, latent embedding visualizations, statistical significance evaluations, and qualitative mineral mapping results.

A.1 EXPERIMENTAL SETUPS

Datasets. We evaluate the proposed framework on **THREE** Martian hyperspectral datasets obtained from **CRISM** (Murchie et al., 2007; Xi et al., 2025b; Kamps et al., 2020), covering diverse geological regions: **Holden Crater** (HC, 418×595 pixels, 440 bands) with evidence of past water activity; **Nili Fossae** (NF, 478×593 pixels, 425 bands), rich in carbonates and hydrothermal alteration products; and **Utopia Planitia** (UP, 478×595 pixels, 432 bands), a northern plain with subtle spectral contrasts. These datasets provide complementary test cases of varying geological complexity.

Experimental Setup. Spectral dimensionality is reduced to 15 using *PCA* (Abdi & Williams, 2010), and hyperspectral patches of size 13×13 are extracted as used in (Xi et al., 2025a). The model is trained with a latent dimension of 64 and a diffusion horizon of $T = 5$ for 100 epochs using Adam (Kingma & Ba, 2015) with a learning rate of 10^{-4} and a batch size of 8. The noise variance sequence $\{\beta_t\}$ is linearly spaced in $\left[10^{-4},\, 2 \times 10^{-2}\right]$. The 3D CNN encoder comprises three layers with 16, 32, and 32 filters, while the class-conditioned latent diffusion network employs two 256-unit dense layers with 64-dimensional time and class embeddings. All hyperparameters are fixed across datasets. 10 LABELED SAMPLES PER CLASS are used for training, with the rest reserved for testing similar to (Xi et al., 2025a).

A.2 EXTENDED METHODOLOGY

Theoretical Formalizations. Let $\mathcal{C}_{\text{shared}} : \mathbb{R}^d \to \Delta^{C-1}$ denote the shared classifier, assumed to be L-Lipschitz continuous, and let $g_\phi(\cdot)$ denote the reverse-time denoising operator. The forward diffusion process induces a Markov chain $\mathbf{z}_0 \to \mathbf{z}_t$ with conditional distribution $q(\mathbf{z}_t \mid \mathbf{z}_0)$, while the reverse diffusion learns an approximation of the conditional expectation

$$\hat{\mathbf{z}}_0 = g_\phi(\mathbf{z}_t, t, y) \;\approx\; \mathbb{E}[\mathbf{z}_0 \mid \mathbf{z}_t, y]\,. \tag{8}$$

Reverse-time classification consistency enforces a bounded expected classification risk over noisy latent variables:

$$\mathbb{E}_{t,\mathbf{z}_t}[\mathcal{L}_{\text{ce}}(\mathcal{C}_{\text{shared}}(\mathbf{z}_t), y)] \le \varepsilon. \tag{9}$$

Here, the expectation $\mathbb{E}_{t,\mathbf{z}_t}[\cdot]$ is taken over the diffusion time index $t \sim \mathcal{U}(\{0, \ldots, T\})$ and the corresponding noisy latent variables $\mathbf{z}_t$ induced by the forward diffusion process, while $\varepsilon > 0$ denotes an upper bound on the expected classification risk enforced by RTCC.

Since the cross-entropy loss upper bounds the 0–1 misclassification loss, Markov's inequality implies that, for any $\delta > 0$,

$$\mathbb{P}_{t,\mathbf{z}_t}(\arg\max \mathcal{C}_{\text{shared}}(\mathbf{z}_t) \neq y) \;\le\; \mathbb{P}(\mathcal{L}_{\text{ce}}(\mathcal{C}_{\text{shared}}(\mathbf{z}_t), y) \ge \delta) \;\le\; \frac{\varepsilon}{\delta}. \tag{10}$$

Hence, with probability at least $1 - \varepsilon/\delta$ over $(t, \mathbf{z}_t)$, the noisy latent $\mathbf{z}_t$ lies in the classifier decision region

$$\mathcal{R}_y = \left\{\mathbf{z} \in \mathbb{R}^d \;\middle|\; \arg\max \mathcal{C}_{\text{shared}}(\mathbf{z}) = y\right\}. \tag{11}$$

We further assume that the classifier admits a positive classification margin $\gamma > 0$ on $\mathbf{z}_t$, defined on the logit outputs as

$$s_y(\mathbf{z}_t) - \max_{k \neq y} s_k(\mathbf{z}_t) \;\ge\; \gamma, \tag{12}$$

where $s_k(\cdot)$ denotes the k-th logit of $\mathcal{C}_{\text{shared}}$. Let $\mathcal{C}_{\text{shared}}$ be L-Lipschitz continuous with respect to the Euclidean norm on the latent space. Then, for any perturbation $\Delta\mathbf{z}$ satisfying

$$\|\Delta\mathbf{z}\|_2 \le \gamma / L, \tag{13}$$

the margin in equation 12 remains strictly positive and the predicted class label is preserved.

Proposition 1 (Transitive Semantic Alignment). Under the constraints in equation 8 and equation 9, and assuming $\mathcal{C}_{\text{shared}}$ is L-Lipschitz continuous, the denoised latent $\hat{\mathbf{z}}_0$ satisfies

$$\|\mathcal{C}_{\text{shared}}(\hat{\mathbf{z}}_0) - \mathcal{C}_{\text{shared}}(\mathbf{z}_t)\|_2 \le L\, \|\hat{\mathbf{z}}_0 - \mathbf{z}_t\|_2\,, \tag{14}$$

and therefore lies, with high probability, within an $O(\delta)$-neighborhood of the decision region $\mathcal{R}_y$, where δ is controlled by the reverse diffusion reconstruction error.

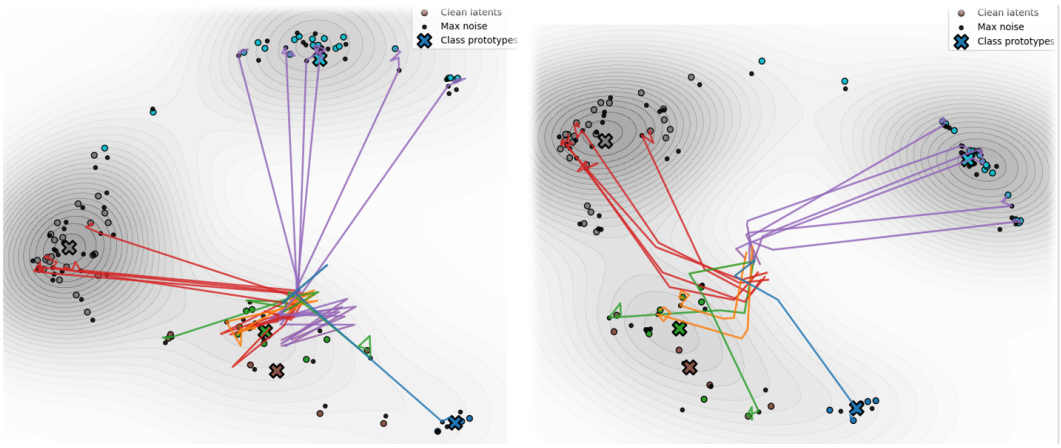

Figure 2: **Latent diffusion geometry and reverse-time flow visualization**. We visualize latent states across noise levels and the corresponding reverse diffusion trajectories. Gray density contours indicate the empirical latent manifold induced by the encoder. Colored points denote clean latents, while black points correspond to maximally noised latents. Polylines depict reverse diffusion trajectories that reflects the direction and relative magnitude of latent updates across timesteps. **Left: (Ours)** without reverse-time consistency (RTCC), reverse diffusion velocities converge toward a class-agnostic region, producing a single dominant manifold basin and frequent cross-class trajectory mixing. **Right: (Ours)** with RTCC, the latent manifold decomposes into class-specific basins of attraction. Reverse diffusion velocities align with manifold gradients and consistently guide samples toward their corresponding class prototypes, yielding smooth and semantically coherent trajectories.

Proof. The bound in equation 14 follows directly from the Lipschitz continuity of $\mathcal{C}_{\text{shared}}$. The reverse diffusion operator g_ϕ is trained to approximate the conditional expectation $\mathbb{E}[\mathbf{z}_0 \mid \mathbf{z}_t, y]$ and minimizes the diffusion reconstruction loss $\mathbb{E}\big[\|\mathbf{z}_0 - \hat{\mathbf{z}}_0\|_2^2\big]$. Since the forward diffusion process generates $\mathbf{z}_t$ via variance-controlled perturbations of $\mathbf{z}_0$, this training objective yields a bounded expected deviation between $\hat{\mathbf{z}}_0$ and $\mathbf{z}_t$ that is controlled by the diffusion noise level. Combined with the margin preservation result from equation 13 and the high-probability correctness guarantee in equation 10, it follows that $\hat{\mathbf{z}}_0$ remains within the classifier decision region $\mathcal{R}_y$ with high probability.

Thus, class-discriminative structure imposed on noisy latent variables propagates transitively to the denoised representation without requiring explicit classification supervision on $\hat{\mathbf{z}}_0$. The Lipschitz continuity of $\mathcal{C}_{\text{shared}}$ constitutes a minimal regularity condition for formalizing semantic invariance under latent perturbations: without it, arbitrarily small changes in latent space could induce unbounded variations in classifier outputs. In our setting, this assumption is mild and well justified, as $\mathcal{C}_{\text{shared}}$ is implemented as a lightweight linear classifier with bounded weights acting on normalized latent representations. Moreover, reverse-time classification consistency enforces output stability across diffusion-induced perturbations, effectively acting as an implicit Lipschitz regularizer during training. □

Reverse-Time Diffusion as a Semantic Flow on the Latent Manifold. Fig. 2 (right) reveals that reverse-time latent diffusion behaves less like stochastic denoising and more like a structured dynamical flow over the learned latent manifold. The gray density contours define an energy landscape induced by the encoder, where high-density regions correspond to stable latent configurations. In this view, each reverse diffusion step applies a velocity update that moves a noisy latent state along the local geometry of this landscape.

While, without reverse-time consistency (left), the diffusion velocity field resembles a radial sink: trajectories originating from diverse regions are pulled toward a single class-agnostic basin. Although denoising is achieved, the absence of semantic constraints causes diffusion flow lines to intersect and overlap in a zig-zag pattern. This behavior suggests that the latent manifold collapses into a dominant attractor, where class information is weakly preserved and reverse diffusion behaves isotropically with respect to semantic structure.

In contrast, our model with RTCC induces a fundamentally different geometry (right). With RTCC, the latent space decomposes into multiple class-specific basins of attraction, each anchored by a learned class prototype. Reverse diffusion velocities now align with manifold gradients that are locally anisotropic, guiding samples toward their corresponding semantic attractors. Rather than collapsing toward global minimum, trajectories follow smooth, directed paths that respect class boundaries. RTCC preserves the global contractive nature of reverse diffusion by inducing a multi-center latent geometry, where each class defines its own low-energy region.

From a physical perspective, RTCC transforms reverse diffusion from a purely noise-driven relaxation process into a semantically conditioned flow field, where class prototypes act as stable fixed points and reverse diffusion integrates a vector field defined over latent space. *This explains how discriminative structure can propagate transitively from noisy latents to clean representations without explicit supervision at the clean state*.

A.3 Extended Results

Scaling Behavior. Fig. 3 analyzes the impact of the embedding dimension d on model complexity as measured by the total number of trainable parameters. For small embedding dimensions ($d \leq 32$), the parameter count increases only marginally, from 110.6K to 127.2K, indicating a near-linear scaling regime. In this range, embedding-dependent components constitute a relatively small portion of the overall architecture, and the dominant parameter contributions arise from d-independent layers. As the embedding dimension increases beyond $d = 32$, the influence of embedding-related layers becomes more pronounced, and leads to a faster growth in the number of parameters. Specifically, the model reaches 148.1K parameters at $d = 64$ and 316.4K at $d = 256$. Despite this acceleration, the growth remains sub-quadratic with respect to d, suggesting that the architecture avoids dense pairwise interactions or fully parameterized transformations that would otherwise induce quadratic scaling. This controlled scaling behavior highlights an important design property of the proposed model: increased representational capacity through higher-dimensional embeddings does not come at the cost of excessive parameter expansion. Consequently, the architecture supports exploration of larger embedding dimensions while maintaining computational efficiency and reducing the risk of over-parameterization, which is essential for stable training and generalization in higher-dimensional latent spaces.

Table 2: Model complexity comparison on the Holden Crater dataset.

Methods	Parameters (K)	Training Time (s)
3DCNN (Li et al., 2017)	132.7	**1.55**
SSRN (Zhong et al., 2018)	135.9	2.20
A^2S^2K (Roy et al., 2021)	**105.3**	2.72
Spectral-F (Hong et al., 2022)	130.1	21.72
GAHT (Mei et al., 2022)	946.8	3.38
SSFTT (Sun et al., 2022)	152.8	1.82
Morph-F (Roy et al., 2023)	141.7	4.19
MCTGCL (Xi et al., 2025a)	247.0	64.53
OURS	148.0	10.29

Ablative Study. Table 3 reports a unified ablation on the Holden Crater dataset, analyzing the effects of loss design, diffusion depth, and latent dimensionality. The full model, jointly optimizing the RTCC loss ($\mathcal{L}_{\mathrm{rtcc}}$) and diffusion regularization ($\mathcal{L}_{\mathrm{diff}}$), achieves the best performance across all metrics, indicating their complementary roles. Removing $\mathcal{L}_{\mathrm{diff}}$ consistently degrades OA and Cohen's κ, confirming that diffusion acts as a structural prior that stabilizes latent geometry. Disabling both terms leads to severe performance collapse, showing that neither objective is sufficient in isolation. Varying the diffusion depth T reveals a non-monotonic trend: shallow diffusion limits semantic propagation, while excessive diffusion oversmooths class structure. An intermediate depth ($T = 5$) provides the best trade-off, suggesting that effective regularization requires controlled stochastic refinement rather than aggressive smoothing. Analysis of the latent dimension d exposes a clear capacity–generalization trade-off. Low-dimensional embeddings underfit, while performance peaks at $d = 64$–128 before declining due to over-parameterization. These findings are consistent with the scaling behavior in Fig. 3 and validate the existence of an optimal latent dimensionality. Based on the ablation results, we recommend using the full loss formulation with both $\mathcal{L}_{\mathrm{rtcc}}$ and $\mathcal{L}_{\mathrm{diff}}$, an intermediate diffusion depth of $T = 5$, and a latent dimension in the range $d = 64$–128.

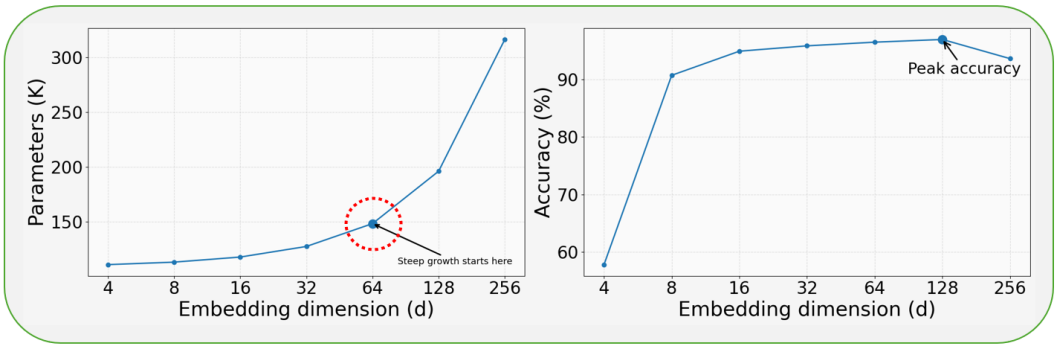

Figure 3: **Effect of embedding dimension** d on **model complexity** and **performance** on HC. **Left:** Parameters scale slowly at low d and faster at high d. **Right:** Accuracy increases with d, peaks, then slightly declines.

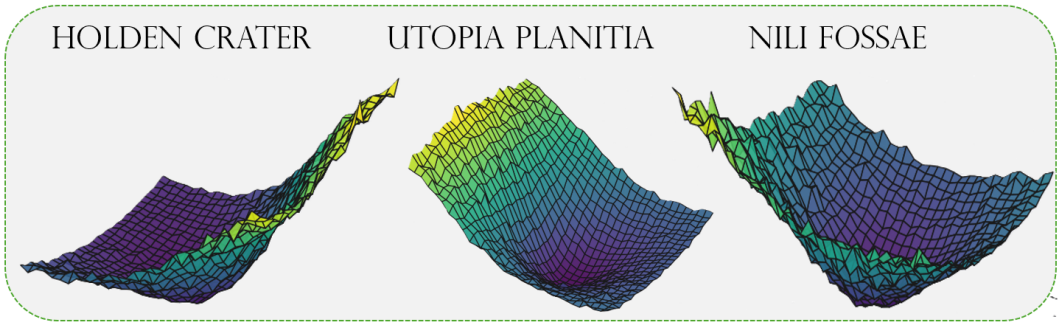

Figure 4: **Loss landscapes over various datasets** obtained by our model.

Loss Landscape Analysis. Fig. 4 depicts empirical loss landscapes of our model on three Martian hyperspectral datasets, evaluated along low-dimensional parameter subspaces around converged solutions. For Holden Crater, the landscape exhibits a steep and well-defined basin, indicating strong gradient alignment and discriminative supervision consistent with its relatively homogeneous spectra. Utopia Planitia shows a smoother, broader minimum, reflecting improved optimization stability and the regularizing effect of latent diffusion. The Nili Fossae landscape is more complex, with curved directions and moderate non-convexity due to higher geological variability, yet remains smooth and navigable. This indicates a balance between model expressiveness and regularization.

Embedding Analysis. Fig. 5 shows t-SNE of the learned latent representations across three Martian terrains. The embeddings display a geometry consistent with neural collapse, where samples within each class form compact clusters and different classes are well separated. This structure, resulting from the interplay of RTCC, which enforces prediction invariance under stochastic perturbations—and latent diffusion—that regularizes class-conditioned trajectories, reduces within-class variance and enlarges inter-class margins. Consequently, the classifier operates on highly structured representations, explaining the strong performance and low variance as seen in Table 4. The consistency across datasets suggests that the latent space approaches a well-regularized geometry.

Statistical Analysis. Table 4 reports a statistical comparison between the proposed method and MCTGCL (Xi et al., 2025a) across three hyperspectral datasets over 10 independent simulations. Across all datasets and evaluation metrics, our method consistently outperforms the baseline, with improvements that are not only substantial in magnitude but also statistically reliable. The observed Cohen's d values exceed 1.5 in nearly all cases and surpass 4.0 on the NF dataset, corresponding to very large to extreme effect sizes under conventional statistical interpretations. Such magnitudes indicate that the performance gains are not attributable to stochastic variation but instead reflect a systematic and practically meaningful advantage of the proposed approach. Moreover, the 95% confidence intervals (CI) associated with our method are consistently narrower than those of MCT-

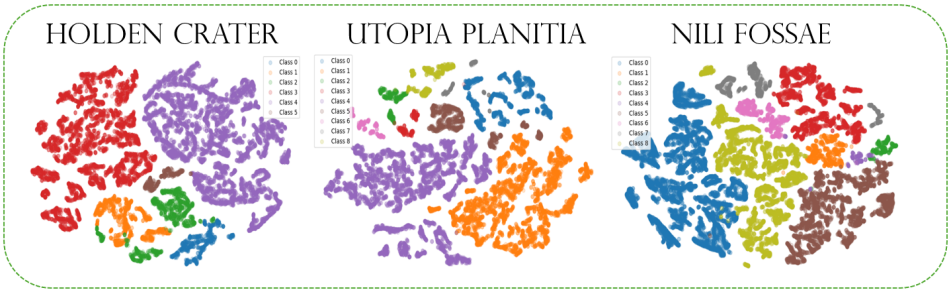

Figure 5: **t-SNE** of latent representations learned on multiple Martian hyperspectral datasets.

Table 3: Ablation study on the HOLDEN CRATER dataset.

Setting	OA (%)	AA (%)	κ
Effect of Loss Functions ($\mathcal{L}$)			
$\mathcal{L}_{\text{rtcc}}$ (✓), $\mathcal{L}_{\text{diff}}$ (✓)	**96.51**	**95.26**	**94.89**
$\mathcal{L}_{\text{rtcc}}$ (✓), $\mathcal{L}_{\text{diff}}$ (×)	96.22	94.65	94.61
$\mathcal{L}_{\text{rtcc}}$ (×), $\mathcal{L}_{\text{diff}}$ (×)	5.620	16.95	0.490
Effect of Diffusion Steps (T)			
1	96.07	94.70	94.26
2	95.76	93.74	93.80
5	**96.51**	**95.26**	**94.89**
10	95.49	92.06	93.42
100	94.50	90.74	91.96
Effect of Latent Dimension (d)			
4	57.77	66.07	44.40
8	90.74	90.07	86.76
16	94.94	93.30	92.61
32	95.87	94.66	93.98
64	96.51	95.26	94.89
128	**96.98**	**95.95**	**95.58**
256	93.64	88.79	90.69

GCL, suggesting reduced performance variance and enhanced robustness across repeated trials. We also observe that, the lower bounds of the confidence intervals for our method remain well above the corresponding mean performance of the baseline in all datasets, providing strong empirical evidence that the observed improvements are statistically significant. This behavior indicates that the proposed diffusion mechanism effectively regularizes latent trajectories, constraining stochastic evolution while preserving class-discriminative structure. Concurrently, the RTCC objective enforces decision-level invariance under stochastic perturbations, leading to stable predictions and improved agreement, as reflected by consistently higher κ values.

Table 4: Statistical comparison between our method and MCTGCL (Xi et al., 2025a) across three hyperspectral datasets over 10 simulations.

Dataset	Metric	OURS	95% CI	MCTGCL Xi et al. (2025a)	d
HC	OA (%)	96.51±0.94	[95.93,97.09]	91.99±2.32	2.55
	AA (%)	95.26±1.01	[94.63,95.89]	93.30±1.40	1.61
	Kappa	0.9489±0.0136	[0.9405,0.9573]	0.8839±0.0326	2.60
NF	OA (%)	92.37±1.08	[91.70,93.04]	81.20±3.65	4.15
	AA (%)	90.08±0.64	[89.68,90.48]	82.68±3.09	3.32
	Kappa	0.9041±0.0134	[0.8958,0.9124]	0.7657±0.0445	4.21
YP	OA (%)	90.85±1.50	[89.92,91.78]	86.93±2.14	2.12
	AA (%)	90.64±0.99	[90.03,91.25]	88.68±1.55	1.50
	Kappa	0.8742±0.0200	[0.8618,0.8866]	0.8219±0.0276	2.14

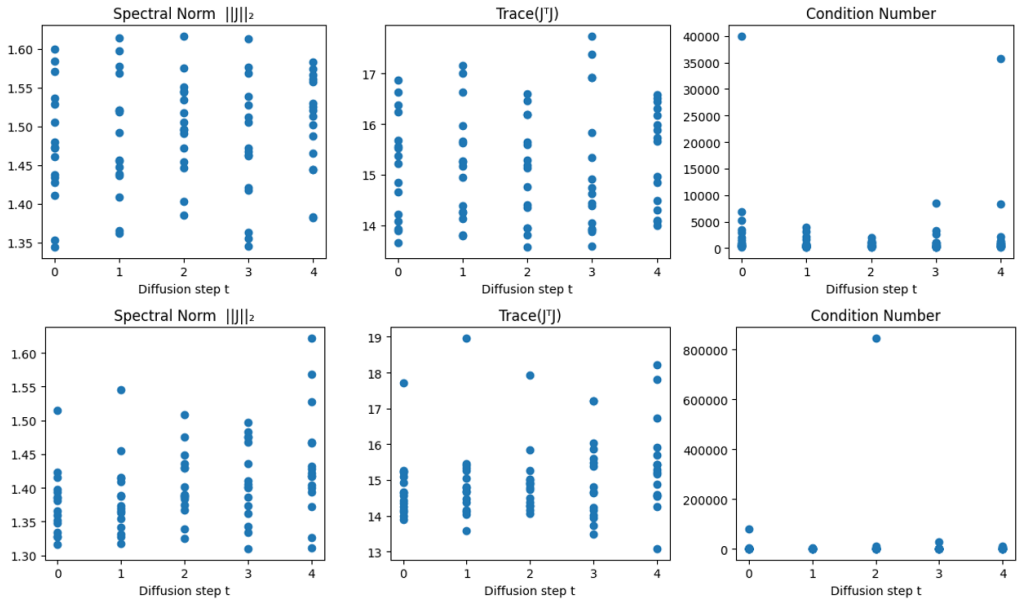

Figure 6: **Jacobian spectrum flow of the reverse diffusion mapping**. Top: our class-conditioned latent diffusion with RTCC. Bottom: the same model without RTCC. RTCC yields bounded spectral norms and stable Frobenius energy while suppressing extreme condition numbers, indicating well-conditioned, class-aligned reverse dynamics. Without RTCC, the reverse mapping exhibits severe anisotropy and near-singular Jacobians.

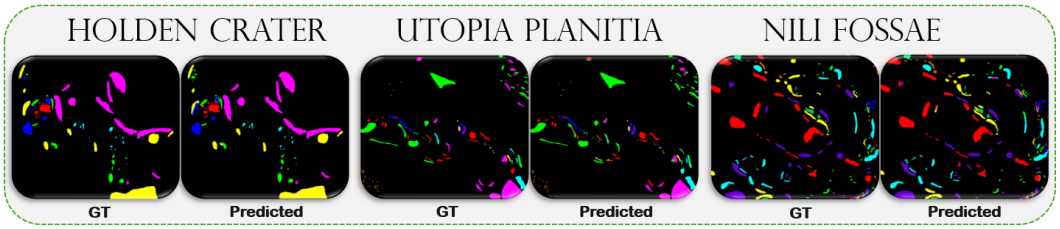

Figure 7: **Mineralogical maps of Mars inferred by our model for three regions of interest:** HOLDEN CRATER, UTOPIA PLANITIA, and NILI FOSSAE.

Latent Geometric Regularization via RTCC. Fig. 6 reveals that RTCC fundamentally alters the geometry of the reverse diffusion dynamics rather than merely improving classification accuracy. While both variants maintain bounded spectral norms, indicating non-explosive mappings, the RTCC-enabled model exhibits substantially more stable Frobenius energy and suppresses extreme Jacobian condition numbers across diffusion steps. This behavior implies that reverse diffusion remains well-conditioned and aligned with class-discriminative directions, even at higher noise levels. In contrast, removing RTCC leads to severe anisotropy and near-singular Jacobians, suggesting that denoising occurs along poorly conditioned latent directions that are decoupled from semantic structure. These results indicate that RTCC acts as an implicit geometric regularizer, stabilizing reverse-time dynamics by enforcing class-consistent latent transformations. The three visualizations report the spectral norm, the trace of $J^\top J$, and the Jacobian condition number of the reverse diffusion mapping across diffusion steps. Together, they characterize the magnitude, overall energy, and numerical conditioning of the reverse-time latent transformation.

Classification Maps. Qualitative visuals in Fig. 7 show that the our method achieves **accurate** mineral mapping on the Mars, and most mineral classes are reliably identified and spatially well *delineated* in the classification maps. Minor misclassifications are observed primarily on small or spatially sparse regions, which can be attributed to limited training samples and spectral similarity with dominant minerals.

Discussions and Limitations. We highlight several practical considerations of the current framework. First, Martian hyperspectral datasets inherently lack direct ground truth, and labels are typically derived from expert spectral interpretation, which may introduce uncertainty. While our method assumes fixed labels, its design—imposing supervision on latent representations and enforcing consistency through diffusion—offers a degree of robustness to such noisy annotations. Nevertheless, explicit modeling of label uncertainty represents an important direction for future work. Second, to ensure a fair comparison with prior studies, models are trained separately on each dataset. Extending the framework to learn a unified model capable of generalizing across heterogeneous Martian regions, where spectral distributions can differ substantially, presents both a significant challenge and an exciting opportunity for future exploration. Finally, given the limited size and nature of Martian hyperspectral data, our focus here is on demonstrating the effectiveness of the proposed method keeping the exhaustive domain-specific analysis limited. As a result, the semantic interpretation of classes (e.g., mineral types) is not explored in depth, leaving a promising avenue for collaboration with domain experts in planetary science.