CROSS-DOMAIN HYPERSPECTRAL IMAGE CAPTIONING: REPRESENTATION AND LANGUAGE COLLAPSE

Srimanreddy Bommireddy and Mrinmoy Ghorai

Indian Institute of Information Technology Sri City, Chittoor, India

ABSTRACT

Hyperspectral image (HSI) captioning maps a spectral image patch to a sentence describing the observed land cover. Although recent datasets support this task, a captioner may encounter scenes unlike those used for training. We examine this problem using a Frequency-Aware Hierarchical Mamba (FAHM) encoder with a Bootstrapping Language-Image Pre-training (BLIP) decoder, transferring between Indian Pines and Kennedy Space Center. The unigram Bilingual Evaluation Understudy (BLEU) score falls from 0.1552 to 0.0310 in one direction and from 0.1441 to 0.0406 in the other. Visual features from the unseen scene remain strongly separated and become more concentrated; the decoder also repeats captions, with duplicate rates of 94.3% and 99.2%. Spectral augmentation, semantic alignment, variance regularization, and unsupervised feature alignment do not consistently restore performance. These findings show that scene changes affect both spectral representations and the language needed to describe them.

***Index Terms*—** domain adaptation, domain shift, hyperspectral imaging, image captioning, vision and language models

1. INTRODUCTION

Hyperspectral images record a spectrum at each pixel instead of the few broad color channels available in ordinary images. These measurements are useful for distinguishing materials and land cover, including vegetation and soil that may look similar in visible imagery. Much of hyperspectral image (HSI) analysis produces a class label or a numerical estimate. Captioning asks for something different: given a small spatial and spectral patch around a pixel, the model generates a natural language description of that location. For example, a caption might describe alternating soil lines and green stalks rather than simply assign the pixel an agricultural class. This could make HSI products easier to inspect and communicate in remote sensing workflows.

Recent HSI foundation models have expanded spectral representation learning [1, 2]. Image captioning models such as Bootstrapping Language-Image Pre-training (BLIP) connect visual features with text generation [3]. HyperCap brings these directions together by providing captions for pixels in four established HSI datasets [4]. Most caption evaluations, however, train and test on different pixels from the same scene. They tell us less about whether a model can describe a new scene with different land cover and acquisition characteristics.

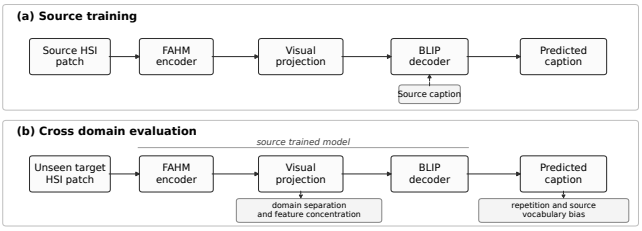


Fig. 1. Cross-domain HSI captioning. A source-trained FAHM+BLIP model receives a target scene with different spectral, semantic, and linguistic statistics. We analyze the resulting feature and language collapse.

That distinction matters for hyperspectral remote sensing. The spectral response of a material varies with the scene and acquisition conditions, and the range of objects being observed can also change. Indian Pines (IP) contains many agricultural categories; Kennedy Space Center (KSC) contains wetlands and coastal vegetation. An IP caption may describe husks or crop residue, whereas KSC references often describe marshes, mud and water. A captioner moved from one dataset to the other must therefore handle changes in the imagery as well as in the concepts and words needed for the caption.

We test this question in both directions using a Frequency-Aware Hierarchical Mamba (FAHM) encoder [5] coupled to BLIP. We first evaluate captioning performance on the new scene, then examine the visual features and generated sentences to understand the failure. Finally, we test spectral perturbation, semantic alignment, feature variance regularization and alignment with unlabeled target images. Our contribution is an empirical diagnosis, not a new captioning architecture: we show that visual features become separated and concentrated while the decoder falls back on repetitive source language, and that simpler feature alignment methods do not consistently correct this behavior.

2. PROTOCOL

2.1. Datasets and preprocessing

We use Indian Pines (IP) and Kennedy Space Center (KSC) from HyperCap [4]. After background removal, IP contains 10,249 captioned pixels across 16 classes and KSC contains 5,211 pixels across 13 classes. Their reference captions use 2,989 and 2,369 distinct word types, with 1,568 occurring in both datasets. IP contains agriculture-oriented terms such as *straw*, *windrows*, *husk*, and *furrow*; KSC contains terms such as *marsh*, *mud*, *silty*, and *wetland*.

IP has 200 available spectral bands and KSC has 176. We linearly interpolate IP spectra to 176 positions so that the same network can process both scenes; this step is called spectral harmonization. Each pixel spectrum is clipped to its 1st and 99th percentiles and standardized, and an 11×11 patch provides the spatial neighborhood around the captioned pixel. We use fixed stratified 70/10/20 training, validation and test splits with seed 42. A separate IP model using the native 200 bands tests whether the interpolation itself explains the drop.

2.2. Captioner and evaluation

The FAHM encoder represents each spectral and spatial patch with a 64-dimensional vector. A learned linear layer maps it to a 768-dimensional visual token that the BLIP text decoder uses when predicting the next words of a caption. Source models use AdamW with learning rate 5×10^{-5}, batch size 16, and early stopping on source validation BLEU-1. Generation uses beam size 4, between 4 and 30 new tokens, a no-repeat trigram constraint, and repetition penalty 1.1. We report BLEU-1 through BLEU-4 scores to measure n-gram overlap [6], Recall-Oriented Understudy for Gisting Evaluation (ROUGE-L) for longest shared word sequences [7], and Consensus-based Image Description Evaluation (CIDEr), which weights agreement by word importance [8].

These are results of a controlled reimplementation, not exact reproductions of the original HyperCap table. We use fixed splits and a documented decoding setup because the original random splits and full evaluation configuration are not available. All comparisons in this study use the same protocol within each evaluation setting.

3. CROSS-DOMAIN FAILURE ANALYSIS

3.1. Caption accuracy

Table 1 shows the main result. IP→IP BLEU-1 is 0.1552 but falls to 0.0310 on KSC, an 80.1% relative decrease. KSC→KSC reaches 0.1441 but falls to 0.0406 on IP, a 71.8% decrease. Higher-order overlap is even more fragile, with cross-domain BLEU-4 equal to zero in both directions.

Table 1. Caption scores within each scene and after zero-shot transfer.

Train	Test	B1	B2	B3	B4	R-L	CIDEr
IP	IP	.1552	.0891	.0614	.0442	.1277	.3459
IP	KSC	.0310	.0035	.0000	.0000	.0309	.0224
KSC	KSC	.1441	.0695	.0393	.0232	.1309	.2608
KSC	IP	.0406	.0070	.0013	.0000	.0371	.0165

Table 2. Representation diagnostics for zero-shot transfer (D/S: centroid distance over average domain spread).

Source	Space	Domain acc.	D/S	Src. cos.	Tgt. cos.
IP	FAHM-64	1.000	1.776	.340	.785
IP	Proj.-768	1.000	1.810	.547	.841
KSC	FAHM-64	1.000	1.618	.289	.859
KSC	Proj.-768	1.000	1.625	.582	.897

A model using all 200 native IP bands reaches BLEU-1 0.1445, while the harmonized 176-band model reaches 0.1552 within IP. The single comparison does not establish a benefit from interpolation, but it makes band matching an unlikely explanation for the much larger transfer gap.

3.2. Representation separation and concentration

We analyze both the 64-dimensional FAHM feature and the 768-dimensional token passed to BLIP. A logistic regression domain classifier distinguishes source from target features with 1.000 accuracy in both spaces and transfer directions. This indicates that dataset identity remains easy to recover from the representation. We also report the distance between source and target feature centroids divided by their average within-domain spread (D/S), where spread is the mean distance of features from their own domain centroid. Projected D/S is 1.810 for IP→KSC and 1.625 for KSC→IP.

Unseen features are also much more similar to one another. We compute mean pairwise cosine similarity across feature vectors; values closer to one indicate vectors that point in similar directions. In the projected space it rises from 0.547 on source IP to 0.841 on KSC, and from 0.582 on source KSC to 0.897 on IP (Table 2). This is evidence of feature concentration, not proof that the features become identical.

3.3. Language collapse and source bias

The representation shift coincides with severe repetition in generated language. IP→KSC produces only 296 distinct captions for 5,211 samples; one sentence accounts for 39.2% of predictions. KSC→IP produces only 82 distinct captions for 10,249 samples, using only 60 distinct words. Exact duplicate rates reach 94.32% and 99.20%, compared with 72.39% and 81.88% on the source scenes. Table 4 shows actual examples from the baseline predictions. The references describe

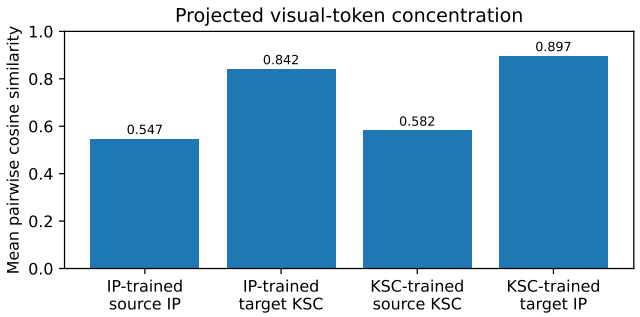


Fig. 2. Mean pairwise cosine similarity of projected visual tokens. Higher values for unseen target scenes indicate greater feature concentration.

Table 3. Caption-collapse statistics.

Direction	Unique	Duplicate	Modal caption	Vocab.
IP→IP	27.61%	72.39%	4.93%	451
IP→KSC	5.68%	94.32%	39.19%	191
KSC→KSC	18.12%	81.88%	7.57%	321
KSC→IP	0.80%	99.20%	11.31%	60

different crop or wetland details, yet the generated captions repeatedly describe familiar source-scene vegetation or water.

We compare unigram distributions using Jensen–Shannon divergence (JSD; lower is closer). IP→KSC predictions are closer to IP references (0.5148) than KSC references (0.6854). KSC→IP predictions are closer to KSC references (0.6206) than IP references (0.7512). KSC-specific terms account for 11.6% of tokens generated by the KSC-trained model on IP even though they occur in 0% of IP reference tokens. Thus the failure is not just low word overlap: generated language retains patterns learned from the source scene.

4. CONTROLLED INTERVENTIONS

The analyses show several effects that occur together, but do not establish that one causes another. We therefore test remedies aimed at different parts of the captioner. Each training objective combines a source caption loss with an optional regularizer:

$$\mathcal{L} = \mathcal{L}_{\text{cap}} + \lambda \mathcal{L}_{\text{reg}}, \tag{1}$$

Here $\mathcal{L}_{\text{cap}}$ measures how well the decoder predicts source captions, and $\mathcal{L}_{\text{reg}}$ is an additional objective whose strength is controlled by λ. The baseline has no additional regularizer.

4.1. Source-only regularization

Spectral randomization makes source spectra less uniform during training by varying gain and bias, adding noise, masking bands, and applying small shifts. IP→KSC BLEU-1 changes only from 0.0310 to 0.0319, while within-domain IP performance declines.

Visual–text alignment encourages a patch embedding to be close to its corresponding source caption embedding and less similar to captions from other patches. We use a symmetric contrastive loss with frozen pooled BLIP text embeddings. It improves IP→KSC BLEU-1 to 0.0428, but hurts KSC→IP from 0.0406 to 0.0248. The geometry moves in the same direction: target cosine decreases from 0.8415 to 0.7842 when transfer improves, but increases from 0.8968 to 0.9396 when transfer worsens. In the latter case, one water-related sentence is emitted for 79.9% of IP samples.

Variance regularization penalizes low variation across the 64-dimensional FAHM features to discourage different patches from receiving overly similar representations. For IP→KSC it lowers projected D/S from 1.810 to 1.508 and target cosine from 0.841 to 0.737, yet BLEU-1 reaches only 0.0373. The features become less concentrated, but the captions do not become correspondingly more accurate.

4.2. Target-aware alignment

Next we test unsupervised domain adaptation (UDA), where unlabeled target training and validation images are available but target captions and labels are not. The target test set stays held out; this is why its no-adaptation IP→KSC BLEU-1 baseline (0.0333) differs from the full-scene zero-shot score in Table 1. Deep Correlation Alignment (CORAL) [9] matches covariance statistics of source and target FAHM features while the BLIP decoder remains frozen. After scaling the small normalized CORAL loss to contribute roughly 3% of the source caption loss initially, BLEU-1 is 0.0326 and CIDEr is 0.0308. Maximum Mean Discrepancy (MMD) [10, 11] instead compares source and target feature distributions through multiple Gaussian kernels. Combining it with a weaker source visual and text alignment loss gives BLEU-1 0.0292. Neither method tells the decoder what an unfamiliar target class should be called, which helps explain why global feature alignment alone is insufficient.

5. DISCUSSION AND CONCLUSION

Our results highlight a practical weakness of HSI captioning across scenes. A model trained on agricultural descriptions is not necessarily able to describe wetlands, and the reverse is equally difficult. In both directions, target features remain distinct from the source distribution and cluster more closely together. The generated captions also become repetitive and lean toward the source vocabulary, as the examples in Table 4 illustrate. These observations occur together, but our analysis does not establish a simple causal link.

The interventions narrow down what may be missing. Spectral augmentation changes the input but does little for transfer. Semantic alignment helps IP→KSC, yet worsens

Table 4. Examples from the zero-shot baseline. Reference captions are paired with generated sentences from the same target patches; no target captions were used for training.

Transfer	Target reference	Generated caption
IP→KSC	Dense irregular patches of muted green and brown hues.	tree shadows create a sense of order across the thick undergrowth.
IP→KSC	Broad mottled textures blending soft earth tones and vegetation.	the ground is covered with thick grass, surrounded by scattered trees.
KSC→IP	Tall stalks stand straight, surrounded by scattered dry organic field debris.	quiet fluid stretches outward, bending light into rhythmic ripples beneath.
KSC→IP	Pale silks peek out slightly from tightly closed greenish husk tips.	quiet fluid stretches outward, bending light into rhythmic ripples beneath.

Table 5. Source-only interventions for IP→KSC.

Method	IP-ID B1	X-domain B1	R-L	CIDEr
Baseline	.1552	.0310	.0309	.0224
Spectral randomization	.1435	.0319	.0298	.0175
Visual–text alignment	.1533	**.0428**	**.0422**	**.0249**
+ variance regularization	.1500	.0373	.0325	.0177

the reverse direction. Variance regularization makes features less concentrated without recovering the same caption improvement, while CORAL and MMD do not substantially improve the held-out KSC score. Matching the distributions of source and target features, $P(Z_s)$ and $P(Z_t)$, does not ensure that the relation between features and words, $P(Y \mid Z)$, is preserved when target concepts are different.

The study is limited to two scenes, one FAHM and BLIP architecture, a fixed split and seed, and target-aware adaptation only from IP to KSC. We have not tested whether training on a larger, more diverse set of HSI scenes would alleviate these failures. That is a useful next experiment, along with evaluation on other sensors and land cover distributions. For hyperspectral remote sensing, robust captioning will require not only features that transfer across acquisition conditions but also language generation that can recognize and describe unfamiliar material and land cover patterns.

6. REFERENCES

[1] D. Hong et al., "SpectralGPT: Spectral Remote Sensing Foundation Model," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 8, pp. 5227–5244, 2024.

[2] D. Wang et al., "HyperSIGMA: Hyperspectral Intelligence Comprehension Foundation Model," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 8, pp. 6427–6444, 2025.

[3] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation," in *Proc. ICML*, 2022, pp. 12888–12900.

[4] A. Das, T. Rachamalla, P. Singh, K. Biswas, V. K. Verma, and S. K. Roy, "Hyperspectral Image Land Cover Captioning Dataset for Vision Language Models," arXiv:2505.12217, 2025.

[5] P. Zhuang et al., "FAHM: Frequency-Aware Hierarchical Mamba for Hyperspectral Image Classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 18, pp. 6299–6313, 2025.

[6] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation," in *Proc. ACL*, 2002, pp. 311–318.

[7] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," in *Text Summarization Branches Out*, 2004, pp. 74–81.

[8] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-Based Image Description Evaluation," in *Proc. CVPR*, 2015, pp. 4566–4575.

[9] B. Sun and K. Saenko, "Deep CORAL: Correlation Alignment for Deep Domain Adaptation," in *Proc. ECCV*, 2016, pp. 443–450.

[10] M. Long, Y. Cao, J. Wang, and M. I. Jordan, "Learning Transferable Features with Deep Adaptation Networks," in *Proc. ICML*, 2015, pp. 97–105.

[11] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A Kernel Two-Sample Test," *J. Mach. Learn. Res.*, vol. 13, pp. 723–773, 2012.