Beyond What's Shared: Recovering Lost Unique Information from Intermediate Layers to Boost Multimodal Geo-Foundation Models

Supplementary Material

A. Additional Ablation Study

Non-linear probing. To show that BWS gains hold beyond linear probing, we repeat the evaluation using MLP probes with official train/val/test data splits [27] (Table 6).

Table 6. BWS benefits non-linear (MLP) probing.

Method	Country (Classification)	Elevation (Regression)
GeoCLIP [67]	0.939	0.869
GeoCLIP + **BWS**	**0.954**	**0.875**
SatCLIP [27]	0.961	0.897
SatCLIP + **BWS**	**0.964**	**0.901**

Controlling for dimensionality. Using GeoCLIP [67], we compress BWS embeddings via PCA to match the final-layer dimension and expand the final-layer embedding to the BWS dimension via random orthogonal transformations, adding no new information. Table 7 shows that compressed BWS outperforms the baseline while orthogonal expansion provides no benefit, indicating the gains are not solely from increased dimensionality.

Table 7. Learned linear probing with dimensionality control.

Method	Avg. Performance
Final layer (baseline)	0.781
BWS (PCA-compressed)	0.808
Final layer (orthogonal-expanded)	0.766
BWS	**0.848**

Layerwise modality-specific (unique) information analysis. We define modality-specific information as information unique to one modality and not accessible from the other. As a proxy for how much each layer retains, we train a linear model to predict modality-specific targets from layer activations: raw coordinates (lon, lat) for the location encoder and per-band spectral means for the image encoder. Using the S2-100k [27] pretraining data, we probe each layer with a linear model and report R^2 (Table 8). Both encoders show a declining trend with depth, further motivating fusing intermediate layers.

Table 8. Unique info recoverability (R^2) decreases with depth.

SatCLIP Encoder	Early Layer	Middle Layer	Final Layer
Location (coord.)	0.827	0.787	0.774
Image (spectral)	0.999	0.917	0.534

Fusing all layerwise information from both encoders. BWS can also be applied to SatCLIP's own vision encoder without external models. Following RANGE [13], we retrieve vision features for downstream coordinates via softmax-weighted similarity between location and image embeddings. Table 9 shows that applying BWS to both encoders yields the strongest results.

Table 9. Applying BWS to both encoders.

Method	Cls. (↑)	Reg. (↑)	Overall (↑)
RANGE	0.921	0.747	0.822
RANGE + BWS (loc.)	0.942	0.752	0.833
RANGE + BWS (both)	**0.950**	**0.790**	**0.859**

B. Broader Applicability of BWS

Generalizing to broader domains. We apply BWS across additional domains and modalities: (1) *medical imaging* with MedCLIP [71] (image-text), (2) *robotics* with TVL [18] (touch-vision-language), and (3) *natural sciences* with AstroM^3 [48] (photometry-spectra-metadata). In each case, BWS fuses intermediate layers within each modality-specific encoder, and the resulting representations are evaluated via linear probing. Table 10 shows consistent gains across domains.

Table 10. BWS benefits diverse contrastive models.

Method	Task	Avg. Performance *Baseline*	Avg. Performance ***+BWS***
MedCLIP	Medical Image-Text Cls.	0.697	**0.717**
TVL	Tactile Attribute Cls.	0.771	**0.797**
AstroM3	Variable Star Cls.	0.820	**0.892**

Diversifying multimodality and downstreams. To test whether BWS generalizes beyond vision-location settings, we pretrain SimCLR [9] with a ResNet50 [21] on paired Sentinel-1 (S1) and Sentinel-2 (S2) imagery from SEN12MS [52]. We evaluate via linear probing on GEO-Bench [31] (six classification, six segmentation tasks). Table 11 shows BWS improves both task types.

Table 11. BWS benefits both GEO-Bench [31] tasks.

Method	Avg. Cls.	Avg. Seg.
SimCLR-RN50 (S1-S2)	0.573	0.312
SimCLR-RN50 (S1-S2) + **BWS**	**0.700**	**0.461**