A Proxy Consistency Loss for Grounded Fusion of Earth Observation and Location Encoders

Supplementary Material

A. Air quality prediction details

A.1. Dataset details

Earth observation features used in the air quality prediction task include time series of: Moderate Resolution Imaging Spectroradiometer (MODIS) Multi-Angle Implementation of Atmospheric Correction (MAIAC) aerosol optical depth (AOD) at 0.47 and 0.55 μm, Daymet meteorology (day length, precipitation, shortwave radiation, min/max temperature, vapor pressure), gridMET near-surface winds (direction and speed), MODIS Normalized Difference Vegetation Index (NDVI), Global Multi-resolution Terrain Elevation Data 2010 (GMTED2010) elevation, and wildfire smoke density from the National Oceanic and Atmospheric Administration (NOAA) Hazard Mapping System (HMS) Smoke product [32].

The reanalysis used for the air quality prediction proxy dataset is generated by assimilating MODIS AOD and Measurement of Pollution in the Troposphere (MOPITT) carbon monoxide (CO) retrievals into the Community Multiscale Air Quality (CMAQ) chemical transport model (CTM), driven offline by Weather Research and Forecasting (WRF) meteorology, using a three-dimensional variational (3D-Var) approach [16].

A.2. Training details

Figure S1b shows the process for generating the spatial train/validation and test splits. To capture variation to the placement of grid boundaries, we repeat this procedure under four checkerboard offsets (original, shifted right by $\delta/2$, shifted up by $\delta/2$, and shifted both up and right by $\delta/2$). In addition, for each offset, we also swap the train/test squares, yielding 8 partitions in total (4 offset $\times$ train/test swap).

Unless otherwise noted, each optimization step samples a minibatch of $B = 256$ labels and an additional minibatch of ρ proxy samples drawn uniformly over space-time (default $\rho = 16$). All models are trained with AdamW [17] for 100 epochs (learning rate 3×10^{-4}; gradient clipping at 1.0) with ReduceLROnPlateau scheduling and early stopping.

For proxy pretraining, we first pretrain only the location-time encoder (together with a proxy-prediction head) for 50 epochs on the proxy prediction task, using the same optimizer and hyperparameter settings as in the main experiments. The Earth observation encoder is not used during this pretraining stage and is initialized randomly for the downstream task. After pretraining, we discard the proxy-prediction head, freeze the pretrained location-time encoder, and train the Earth observation encoder together with the downstream prediction head for 100 epochs. Thus, all methods that use a frozen location encoder (including GeoCLIP, Climplicit, and proxy pretraining) receive the same 100-epoch training. We include this baseline to isolate the value of task-aligned proxy supervision without joint end-to-end fusion.

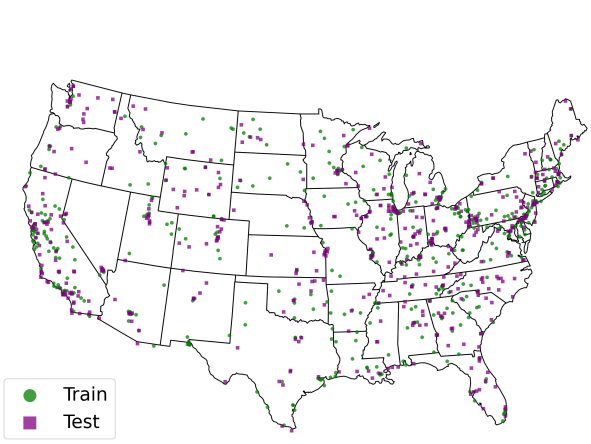
B. Poverty mapping task details

B.1. Training details

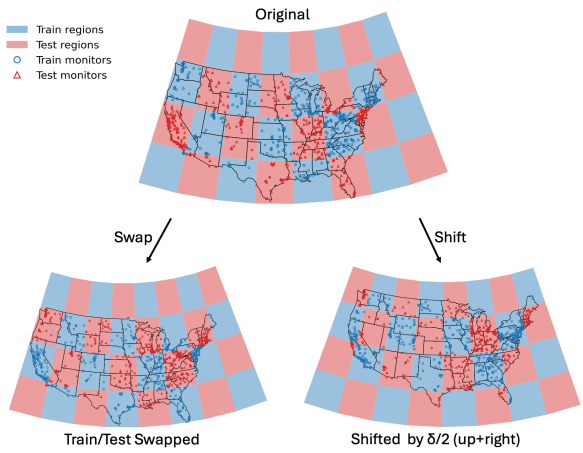
Unless otherwise specified, we train each model for 50 epochs with a learning rate of 0.0001 and a batch size of 128, using an AdamW optimizer [17]. Aside from the loading of pre-trained weights outlined above, all modules are trained simultaneously. For models with a proxy task, we use multiple optimizers to propagate loss through the appropriate parameters.

C. Detailed air quality prediction results

Table S1 supplements the results of Table 1 in the main text.



(a) UAR 50/50 split for $PM_{2.5}$ prediction task.



(b) Systematic checkerboard split ($\delta = 8°$) and augmentations used to measure variation in results.

Figure S1. **Geographic in-sample and out-of-sample evaluation protocols for the $PM_{2.5}$ prediction task.** Plotted points represent EPA stations where in-situ $PM_{2.5}$ measurements are taken. (a) Uniform-at-random (UAR) 50/50 location split, where test sites may be geographically close to training sites. (b) Systematic checkerboard split for geographic extrapolation. We evaluate 8 partitions per δ (4 checkerboard offsets $\times$ train/test swap).

Method	**UAR 50/50**	**Checkerboard** ($\boldsymbol{\delta = 8°}$)	**Time/epoch** (s)
	MAE ($\mu g/m^3$)	**MAE** ($\mu g/m^3$)	
Baselines			
Proxy-only regression	3.174	3.200 ± 0.041	–
Obs. encoder only	2.428 ± 0.003	3.087 ± 0.074	15.85
Proxy-stacked obs. encoder	2.388 ± 0.005	2.974 ± 0.074	16.25
Frozen location-encoder fusion			
GeoCLIP	2.360 ± 0.003	*2.923* ± 0.068	17.55
Climplicit	2.286 ± 0.003	3.062 ± 0.072	17.52
Proxy pretraining (ours)	*2.190* ± 0.006	2.997 ± 0.075	–
Trained location-encoder fusion			
without PCL	2.239 ± 0.005	2.991 ± 0.089	18.40
PCL ($\lambda = 0.2$, $\rho = 16$)	$\mathbf{2.186} \pm 0.003$	$\mathbf{2.857} \pm 0.078$	21.29

Table S1. **Additional metrics and efficiency (MAE and time/epoch) for air quality task.** MAE is reported as mean $\pm$ SE over 5 random seeds for UAR and over 8 partitions (4 offsets $\times$ train/test swap) for checkerboard. Time/epoch is measured on the same hardware as the main experiments and is not reported for proxy-pretrained fusion due to two-stage training. Corresponding R^2, RMSE, and MBE are reported in Table 1.

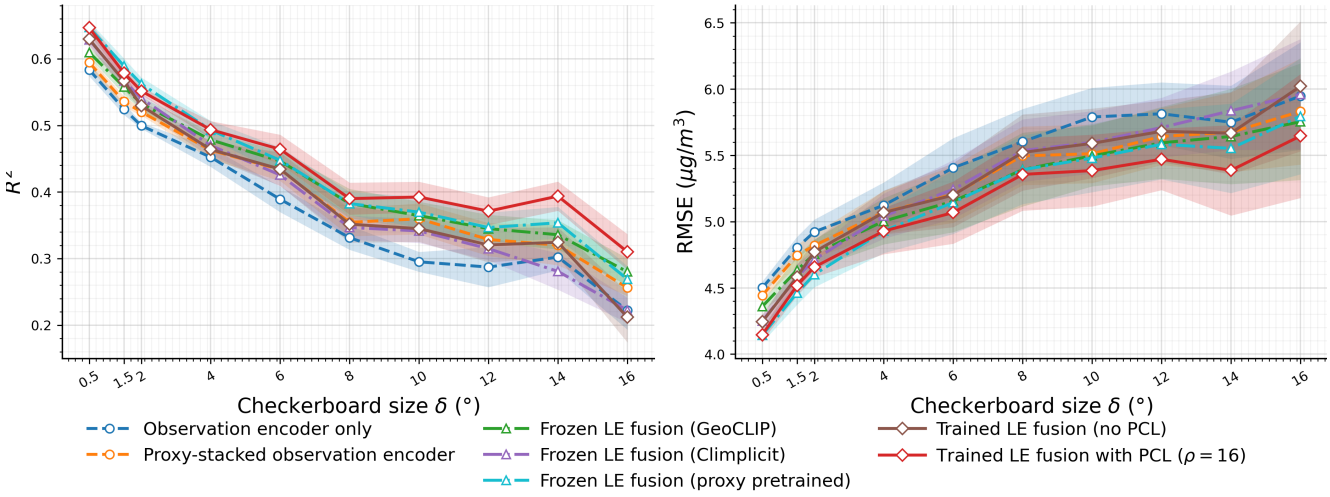


Figure S2. **Air quality prediction performance across checkerboard sizes** δ (°). *Left:* R^2 (higher is better). *Right:* RMSE (lower is better). As δ increases (harder spatial extrapolation), performance generally degrades (decreasing R^2 and increasing RMSE) across all methods. Curves show the mean over 8 checkerboard partitions (4 spatial offsets $\times$ train/test swap), and shaded regions denote ± 1 SE across partitions.