A. Data Sources

Here, we provide a summary of data sources and acquisition. Details are in the supplement. Table 3 lists all data sources.

Satellite Imagery. We leverage aerial imagery from the National Agriculture Imagery Program (NAIP), accessed via the Microsoft Planetary Computer. For each geolocated CAFO point, the nearest cloud-free NAIP image is queried and downloaded using the STAC API. For this experimentation, we collected most recent (2023) data for the studied states.

CAFO datasets. The references to all data sources are in Table 3. We obtained data from multiple sources. Department of Geographical and Sustainability Sciences of IOWA (denoted as IOWA-CAFO Inventory) and animal feeding operation reports from states where available (denoted as State-CAFO Inventory). IOWA-CAFO inventory aggregates CAFO facility data from state environmental agencies across nine southeastern US states. Data is collected from permit databases, nutrient management plans, and agency inspections. Each record includes geolocation, animal type (poultry, swine, beef, dairy), and manure management details. Several US states publish CAFO reports curated by official state agencies using permit records, inspection data, and self-reported nutrient management plans. See Table 3 for the states and the sources (row 3).

Land Use Masks. We utilize national-scale raster products to identify and contextualize agricultural areas. The MRLC National Land Cover Database (NLCD) offers 30m-resolution land cover classifications across 16 categories, including cultivated cropland, grassland, barren land, and pasture (see Table 3). This dataset is further used to generate stratified negative samples based on land cover types and spatial extents.

Table 3. Summary of Core CAFO Data Sources Used in Livestock Detection.

Source	**Description**	**Use**
NAIP Imagery (2023) [52]	High-resolution aerial imagery from the USDA National Agriculture Imagery Program	Visual input for CAFO patch extraction and model training
CAFOMaps [7]	Multi-state (i.e., Alabama, Arkansas, Florida, Georgia, Louisiana, Mississippi, North Carolina, South Carolina, and Texas) labeled dataset containing 6,604 CAFOs with animal type annotations (e.g., poultry, swine, beef, dairy), curated by IOWA researchers	Ground-truth labels for training and validation of CAFO classification models
State-CAFO Inventory [6, 20, 30,35,37,38,41]	Six independent data sources from official state-level CAFO registries corresponding to Indiana, Iowa, Maryland, Michigan, Minnesota, New York, and Delaware, curated from permit records, inspections, and nutrient management plans	Ground truth labels for ML-ready dataset
Land Use Masks (NLCD) [33]	National Land Cover Database (NLCD) used for masking agricultural zones	Used to create negative samples