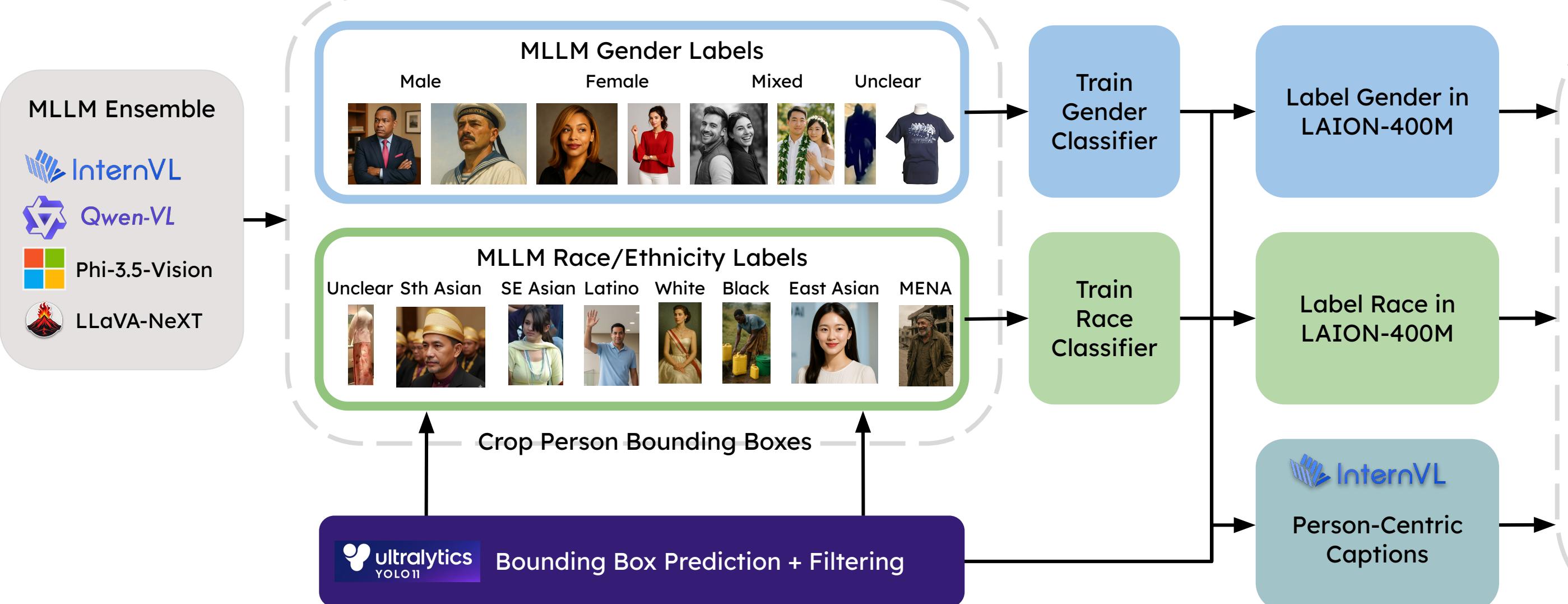# Person-Centric Annotations of LAION-400M

Auditing Dataset Bias and its Transfer to Models

MLLM Ensemble
- InternVL
- Qwen-VL
- Phi-3.5-Vision
- LLaVA-NeXT

MLLM Gender Labels
Male   Female   Mixed   Unclear

MLLM Race/Ethnicity Labels
Unclear   Sth Asian   SE Asian   Latino   White   Black   East Asian   MENA

Crop Person Bounding Boxes

ultralytics YOLO11 — Bounding Box Prediction + Filtering

Train Gender Classifier → Label Gender in LAION-400M

Train Race Classifier → Label Race in LAION-400M

InternVL — Person-Centric Captions

Harmful Correlations
Gender
Crime Words

Dataset Composition
Gender
Male 42.3% 34.1% | Female 35.3% 32.7% | Mixed 6.1% 17.6% | Unclear 16.3% 15.6%

Race / Ethnicity
Black, East Asian, Latino, Middle Eastern, South Asian, Southeast Asian, White
Sentiment

Data-to-Model Bias Transfer
ρ: 0.84, R²: 0.71
CLIP

ρ: 0.80, R²: 0.63
Stable-Diffusion

SAE Identity Representations
caption emb. → SAE →

	Black
Male	basketball
	football
	assault/injury
	championship
	population growth
Female	sports/gymnastics
	consumer confidence
	poverty/hunger
	skin/beauty
	cultural policy