Literature Review & Research Plan

Experimental Physics of Photons at the CMS Detector

Reconstruction, Machine Learning, and Foundation-Model Prospects

Aritra Bal

Experimental Particle Physics (ETP) & Theoretical Particle Physics (ITP)
Karlsruhe Institute of Technology (KIT)

July 2, 2026

This report surveys photon reconstruction and identification at the CMS experiment, reviews the current use of machine learning across the photon reconstruction chain, and provides an original gap analysis and research plan for foundation-model approaches to photon physics. Sections 4–6 and 8.1 are verified against public CMS results and the arXiv literature; Section 7 constitutes original synthesis by the author.

Contents

Executive Summary

- At CMS, a photon is reconstructed as an isolated electromagnetic (EM) shower in the lead-tungstate ($PbWO_4$) ECAL with no matched inner track (unless it converts). The two dominant experimental challenges are photon energy resolution and rejection of $\pi^0 \to \gamma\gamma$ fakes; both set the sensitivity of the flagship $H \to \gamma\gamma$ and $HH \to bb\gamma\gamma$ analyses.
- The current CMS production chain is built on classical multivariate methods (Mustache/DeepSC superclustering, BDT energy regression, BDT photon identification, BDT diphoton vertex and mass-resolution estimators), with deep learning entering at the clustering (DeepSC graph network) and end-to-end image (CNN) stages. No dedicated photon- or ECAL-specific foundation model exists today.
- The genuine open opportunity is a self-supervised, pretrained representation of ECAL crystal-level showers that transfers across photon energy regression, π^0/γ discrimination, conversion classification, and diphoton vertex pointing. Adjacent building blocks exist (jet foundation models, generative calorimeter models with demonstrated cross-geometry transfer), but none has been assembled into a photon representation learner. This report specifies the tasks, datasets, and evaluation metrics needed to build one.

1 The Photon: Basics and Detector Manifestation

The photon (γ) is the gauge boson of the electromagnetic interaction in the Standard Model (SM). It is massless, electrically neutral, and has spin 1 with two physical (transverse) polarization states. Because it carries no charge, it does not ionize matter directly and leaves no track in a tracking detector; it is detected only through the secondary charged particles produced when it interacts with material.

1.1 Photon interactions with matter at LHC energies

Three photon–matter processes compete, with relative importance set by photon energy and the atomic number Z of the absorbing material:

- **Pair production** ($\gamma \to e^+e^-$), dominant at LHC photon energies (tens of GeV to TeV) in the field of a nucleus;
- **Compton scattering**, important at intermediate energies (hundreds of keV to a few MeV);
- **Photoelectric effect**, dominant only at low energies (keV scale).

At the multi-GeV energies relevant to LHC hard-scattering photons, pair production dominates overwhelmingly. In a dense, high-Z absorber, pair production initiates an *electromagnetic shower*: the e^+e^- pair radiates bremsstrahlung photons, which pair-produce again, cascading until the average particle energy falls below the material's critical energy. A calorimeter measures a photon's energy by containing this cascade and collecting the resulting scintillation or ionization signal.

1.2 The CMS photon signature

A prompt, high-energy photon manifests in CMS as:

1. no charged-particle track in the silicon tracker pointing to the energy deposit (unless the photon converted);
2. a compact cluster of energy in the ECAL with a shower shape characteristic of an EM cascade;
3. negligible energy in the hadron calorimeter (HCAL) behind it, quantified by the ratio H/E of hadronic to electromagnetic energy;
4. isolation: little additional activity (tracks, calorimeter energy) in a cone around the candidate.

1.3 Photon conversions

A non-trivial fraction of photons convert to e^+e^- in tracker material before reaching the ECAL. The CMS tracker material budget rises from a minimum of $0.4\,X_0$ at $|\eta| \approx 0$ to $\approx 2.0\,X_0$ at $|\eta| \approx 1.4$, then decreases to $\approx 1.3\,X_0$ at $|\eta| \approx 2.5$, where X_0 denotes a radiation length. A converted photon produces two nearby tracks and a displaced conversion vertex; its ECAL energy is spread in azimuth (ϕ) because the 3.8 T magnetic field bends the e^+ and e^- in opposite directions.

Conversions are simultaneously a complication and a diagnostic handle:

- *Complication:* the same physical photon can be reconstructed through two paths (a pure-ECAL object or a track-seeded object) and must not be double-counted; converted photons have degraded energy resolution.
- *Handle:* conversion tracks point back to the production vertex, valuable for diphoton vertex assignment (Section 3.7), and the measured conversion rate maps the tracker material budget *in situ*.

1.4 CMS ECAL technology

The CMS ECAL is a homogeneous, hermetic scintillating lead-tungstate ($PbWO_4$) crystal calorimeter comprising:

- a **barrel** (EB) of 61,200 crystals covering $|\eta| < 1.479$;
- two **endcaps** (EE) of 7,324 crystals each, covering $1.479 < |\eta| < 3.0$;
- a silicon-strip **preshower detector** (ES) in front of each endcap, covering $1.653 < |\eta| < 2.6$, consisting of two silicon planes interleaved with three radiation lengths of lead.

$PbWO_4$ is a fast scintillator peaking near 425 nm, with approximately 80% of the light emitted within the 25 ns LHC bunch-crossing interval — a property central to the ECAL's timing capability. Scintillation light is read out by avalanche photodiodes (APDs) in the barrel and vacuum phototriodes (VPTs) in the endcaps.

The preshower's fine granularity helps separate a single prompt photon from the closely spaced photon pair of a high-energy $\pi^0 \to \gamma\gamma$ decay (Section 3.1). The barrel achieves better energy resolution than the endcaps owing to less upstream material, lower radiation dose, and finer effective segmentation relative to shower size. For electrons from Z decays reaching the ECAL without significant bremsstrahlung, the measured energy resolution is better than 1.8%, 3.0%, and 4.5% in the pseudorapidity intervals $[0.0, 0.8]$, $[0.8, 1.5]$, and $[1.5, 2.5]$ respectively.

1.5 Electron/photon reconstruction overlap and superclusters

Electrons and photons both begin as ECAL **superclusters** (SCs): groups of neighboring ECAL clusters combined to recover the energy of a single incident particle spread out before or within the ECAL. Spreading occurs because electrons emit bremsstrahlung and photons convert in the tracker; the 3.8 T solenoidal field sweeps this low-energy secondary energy predominantly along ϕ, giving the deposit a characteristic elongated "mustache" shape in the $\Delta\eta$–$\Delta\phi$ plane. The supercluster algorithm collects this ϕ-spread energy back into one object.

Electrons and photons are distinguished downstream by the presence or absence of a matched **Gaussian-Sum-Filter (GSF) track** — a specialized tracking algorithm accounting for bremsstrahlung energy loss. An SC with a matched GSF track is an electron candidate; an SC without one is a photon candidate. This shared starting point is why CMS treats electron and photon ("EGamma") reconstruction as one coupled problem.

2 Why Photon Physics Matters

2.1 $H \to \gamma\gamma$

The decay of the 125 GeV Higgs boson to two photons has a SM branching fraction of only $\approx 0.23\%$, yet it was one of the two channels establishing the boson's discovery in 2012 and remains a leading precision channel. The reason: both decay products are fully measured in the calorimeter with no missing energy, so the diphoton invariant mass $m_{\gamma\gamma}$ reconstructs as a narrow peak on a smoothly falling background. The typical diphoton mass resolution is 1–2%, sufficient to resolve the Higgs peak and measure its mass, couplings, and differential/fiducial cross-sections.

> *The Run 2 (137 fb^{-1}) $H \to \gamma\gamma$ measurement determined the total signal strength to be* 1.12 ± 0.09 *relative to the SM prediction. The Run 3 (2022, 34.7 fb^{-1}) measurement found the inclusive fiducial cross section* $\sigma_{\text{fid}} = 74 \pm 11\,(\textit{stat})\,^{+5}_{-4}\,(\textit{syst})$ *fb, versus the SM prediction of* 67.8 ± 3.8 *fb.*

2.2 $HH \to b\bar{b}\gamma\gamma$

Higgs pair production probes the Higgs trilinear self-coupling κ_λ, fixing the shape of the Higgs potential. The $b\bar{b}\gamma\gamma$ final state has a small combined branching fraction, but it is one of the most sensitive HH channels: the diphoton system gives a clean trigger and narrow mass peak that suppress the enormous QCD multijet background afflicting the higher-branching $b\bar{b}b\bar{b}$ channel, while the $b\bar{b}$ pair supplies most of the rate.

The CMS Run 2 (137 fb^{-1}) search set an observed (expected) 95% CL upper limit of $0.67\ (0.45)$ fb on $\sigma(HH) \cdot \mathcal{B}(\gamma\gamma b\bar{b})$, corresponding to $7.7\ (5.2)$ times the SM prediction, and constrained the self-coupling modifier to $-3.3 < \kappa_\lambda < 8.5$.

2.3 Other photon-involving analyses

- SM diphoton continuum measurements (irreducible $H \to \gamma\gamma$ background and a QCD test in its own right);
- photon+jet production (probe of parton distribution functions and direct-photon QCD dynamics);
- $Z\gamma$ production and anomalous triple-gauge-coupling searches;
- BSM searches: large extra dimensions (ADD) with photon+missing-energy signatures, dark-photon searches, generic photon+MET signatures, and high-mass diphoton resonance searches.

2.4 The general limiting principle

Any analysis using photons is fundamentally limited by two factors: the **photon energy resolution** (setting the width of any mass peak, and hence signal significance) and the **rejection power against** $\pi^0 \to \gamma\gamma$ **fakes** produced in jet fragmentation, misreconstructed as single photons when the two decay photons merge in the calorimeter. Improving either limit improves essentially every photon analysis simultaneously, which is precisely why a shared, high-quality photon representation (Section 6) is attractive.

3 Challenges in Photon Reconstruction and Identification

3.1 π^0/γ discrimination

This is the central identification challenge. A high-p_T π^0 decays to two photons whose opening angle shrinks with energy; above roughly tens of GeV, the two photons deposit energy in overlapping ECAL crystals and form a single cluster nearly indistinguishable from a prompt photon. Discrimination exploits the shower shape: a genuine single photon produces a narrower, more concentrated deposit than a merged

pair. Two variables dominate:

R_9 the ratio of the energy in the central 3×3 crystal array to the total supercluster energy,

$$R_9 = \frac{E_{3\times3}}{E_{\text{SC}}}.$$

A value near 1 indicates a compact, unconverted shower; lower values indicate a spread-out shower (conversion, or a merged/wide deposit). R_9 also separates converted from unconverted photons.

$\sigma_{i\eta i\eta}$ the log-energy-weighted lateral spread of the shower along η within the 5×5 crystal matrix around the seed crystal,

$$\sigma_{i\eta i\eta} = \sqrt{\frac{\sum_i w_i(\eta_i - \bar{\eta})^2}{\sum_i w_i}}, \qquad w_i = \max\left(0,\ 4.7 + \ln\frac{E_i}{E_{5\times5}}\right).$$

A wider deposit (larger $\sigma_{i\eta i\eta}$) is more jet/π^0-like. A companion variable $\sigma_{i\phi i\phi}$ measures the ϕ spread, though the magnetic field reduces its discriminating power.

These, together with H/E, isolation sums, and preshower variables, form the classical identification inputs (Section 4.3).

3.2 Energy resolution and calibration

Achieving 1–2% mass resolution requires intercalibrating tens of thousands of individual crystals, correcting for light-yield non-uniformity, and correcting for shower energy lost to material upstream of the ECAL (bremsstrahlung and conversions). Over the LHC lifetime the crystals suffer radiation damage that reduces transparency; CMS corrects this crystal-by-crystal using a dedicated **laser monitoring system** that continuously measures each crystal's response and applies time-dependent transparency corrections. Intercalibration additionally uses ϕ-symmetry of energy flow, $\pi^0/\eta \to \gamma\gamma$ mass peaks, and $Z \to ee$ events.

3.3 Pileup

Additional pp interactions in the same or nearby bunch crossings (pileup, PU) deposit extra energy that raises fake-photon rates and degrades isolation-based discrimination. The average PU during Run 2 was 27, 38, and 37 for 2016, 2017, and 2018 respectively. CMS mitigates this in isolation sums using the median energy density per unit area, ρ, and **effective-area corrections**: the estimated PU contribution, proportional to ρ, is subtracted from each isolation cone.

3.4 Conversions

Converted photons must be reconstructed without double-counting against the pure-ECAL path and without efficiency loss; the energy regression must additionally correct for the ϕ-spread of conversion energy.

3.5 Endcap-specific challenges

The endcaps face higher PU density, coarser preshower granularity, a larger material budget (more conversions, worse containment), and higher radiation dose, all of which degrade resolution and π^0/γ separation relative to the barrel.

3.6 Spike and anomalous-signal rejection

A detector-specific pathology in the EB is direct ionization of the APDs by highly ionizing particles, producing isolated single-crystal deposits with large apparent energy ("spikes") that mimic high-energy photons. Untreated, spikes would dominate the EM trigger rate: on average one spike with $E_T > 3\,\text{GeV}$

appears per 370 minimum-bias triggers at $\sqrt{s} = 7$ TeV, and up to 60% of EM trigger candidates above a 12 GeV threshold would be spikes if unfiltered.

Because a real EM shower spreads over several crystals (up to $\approx 80\%$ in the central crystal, most of the remainder in the four neighbors) while a spike is confined to one crystal, the topological **Swiss-cross** variable

$$S = 1 - \frac{E_4}{E_1}$$

(where E_1 is the central crystal energy and E_4 the sum of the four adjacent crystals) discriminates spikes from genuine showers, complemented by precise ECAL timing, since spikes and related backgrounds show a broad time distribution relative to the fast ($\approx 80\%$ within 25 ns) $PbWO_4$ scintillation.

3.7 Isolation variable construction

Isolation is built from particle-flow (PF) candidates in a cone (typically $\Delta R < 0.3$) around the photon, summed separately into charged-hadron, neutral-hadron, and photon components, with candidates overlapping the photon's own shower removed ("footprint removal"). A larger cone captures more of the genuine surrounding activity but is more PU-sensitive; a smaller cone is PU-robust but less discriminating. The ρ-effective-area correction restores PU robustness.

3.8 Vertex assignment for diphoton events

Because photons leave no track, they do not directly point to a primary vertex. Identifying the correct vertex matters: if the reconstructed vertex lies within ~ 1 cm along z of the true diphoton vertex, the angular contribution to $m_{\gamma\gamma}$ is subdominant to the energy resolution; otherwise mass resolution degrades substantially. CMS addresses this with a BDT using observables from tracks recoiling against the diphoton system, supplemented by conversion-track pointing when available. The efficiency of assigning a vertex within 1 cm of the true vertex is approximately 79% in Run 2 gluon-fusion Higgs events.

4 State-of-the-Art Reconstruction and Identification Techniques

4.1 Supercluster reconstruction pipeline

The Run 2 production algorithm is the geometrical **"Mustache" superclustering**: starting from a seed cluster above threshold, nearby clusters are added if they fall within a $\Delta\eta$–$\Delta\phi$ mustache-shaped window (bounded by parabolas parameterized by seed η and cluster energy) tuned to contain about 98% of the EM shower energy. Mustache is highly efficient but purely geometric, and thus admits PU and noise contamination; its performance is expected to degrade further in Run 3 as PU and ECAL noise (from aging) rise.

CMS has therefore developed **DeepSC (Deep SuperCluster)**, a graph-neural-network plus self-attention model that (i) decides which clusters to associate to the seed, (ii) regresses an energy correction for PU and leakage, and (iii) classifies the particle type (electron/photon/jet). DeepSC builds a graph over windows around each seed cluster with $p_T > 1$ GeV and runs a TensorFlow model over the resulting structure. It improves energy resolution, particularly in the high-material region $1.0 < |\eta| < 1.5$, and is more robust to PU and noise than Mustache.

4.2 Photon energy regression

After clustering, the raw supercluster energy is corrected to the true photon energy with a multivariate regression: historically, a semi-parametric BDT trained on simulation, taking shower-shape variables, preshower information, local cluster geometry, and ρ as inputs, and predicting both an energy correction and a per-photon energy uncertainty σ_E. This per-photon σ_E is a direct input to the per-event mass-resolution estimate (Section 5.4).

4.3 Photon identification

The production discriminant is an MVA (BDT, implemented in TMVA) combining shower-shape and isolation variables with PU/kinematic terms (ρ, η, uncorrected SC energy) into a single score. Both cut-based and MVA identification are provided, with loose/medium/tight working points; the Run 2 EGamma performance paper defines identification working points at 70%, 80%, and 90% selection efficiency.

4.4 Diphoton vertex selection and mass resolution

Two BDTs work in tandem:

- the **vertex-identification BDT** selects the diphoton vertex using recoil-track and conversion-track observables (average vertex efficiency 75–80% across Run 2 years);
- the **vertex-probability BDT** estimates the per-event probability that the chosen vertex lies within 1 cm of the true one, using inputs including the three highest vertex-ID BDT scores, the number of vertices, $p_T(\gamma\gamma)$, distances to the second- and third-best vertices, and the number of conversion-associated photons.

The per-event mass resolution estimator combines the two per-photon regression uncertainties,

$$\frac{\sigma_m}{m} = \frac{1}{2}\sqrt{\left(\frac{\sigma_{E_1}}{E_1}\right)^2 + \left(\frac{\sigma_{E_2}}{E_2}\right)^2},$$

and events are categorized by σ_m/m to sharpen the final fit.

4.5 Energy scale and resolution calibration strategy

Because real photons have no track, the absolute energy scale is transferred from electrons, which provide a track+cluster cross-check unavailable for photons: $Z \to ee$ events calibrate the scale and, via a Gaussian smearing added to simulation to match the observed $Z \to ee$ mass width, the resolution. Radiative-Z ($Z \to \ell\ell\gamma$) events provide a direct photon-calibration handle, and the laser-monitoring/transparency-correction system maintains stability against crystal aging over the run.

5 Machine Learning in Photon Physics: A Dedicated Deep-Dive

Machine learning is now present across the photon reconstruction chain, at two conceptually distinct levels: **object-level ML** (reconstructing or identifying a single photon from detector information) and **event/analysis-level ML** (combining reconstructed objects to classify signal versus background in a specific analysis). Conflating the two obscures where reconstruction ends and analysis begins; the distinction is maintained explicitly throughout this section.

5.1 Photon energy regression

The classical semi-parametric BDT regression predicting energy and per-photon uncertainty (Section 4.2) remains the baseline. DNN successors and the DeepSC energy-correction head represent the shift toward deeper models; DeepSC retrains dedicated energy corrections and improves resolution over Mustache, especially in high-material regions. The regression's ability to output a calibrated per-photon uncertainty is essential, since that uncertainty propagates directly into σ_m/m.

5.2 Photon identification / π^0 rejection

The production identification discriminant is a BDT on high-level variables. The deep-learning direction instead operates directly on ECAL crystal-level images rather than summarized shower shapes. The pioneering demonstration is an end-to-end CNN trained on 2012 CMS Open Data, which classifies $H \to \gamma\gamma$

signal against backgrounds directly from ECAL crystal energies and can discriminate electron- from photon-induced showers even when the underlying particles are not fully resolved, outperforming purely kinematic classifiers. CMS has since published an end-to-end deep-learning reconstruction of decays to merged photons, using a 32×32 matrix of ECAL crystals around the seed (each pixel encoding one crystal's energy) with a "domain continuation" technique to bridge simulation and data, validated on $\pi^0 \to \gamma\gamma$ decays in collision data. These crystal-image approaches represent the natural high-resolution route to π^0/γ separation but remain supervised, trained from scratch.

5.3 Vertex identification for diphoton systems

The vertex-ID and vertex-probability BDTs (Section 4.4) are the current production ML tools; their inputs are recoil-track kinematics, conversion pointing, vertex multiplicity, and $p_T(\gamma\gamma)$. They feed mass resolution directly through event categorization.

5.4 Per-event mass resolution estimation

The σ_m/m estimator combines the per-photon regression uncertainties with the vertex probability; a dedicated diphoton BDT further predicts the per-event relative mass resolution under correct- and incorrect-vertex hypotheses. This per-event resolution prediction, and its use to weight events in the final fit, is a distinctive CMS technique.

5.5 Conversion reconstruction and classification

Conversion finding uses track-pair and vertex algorithms; R_9 provides a fast converted/unconverted split. No dedicated production-level deep classifier for conversions is documented in the sources reviewed here; this is an area where ML use appears limited beyond the classical approach **[unverified]**.

5.6 Pileup mitigation for isolation and identification

The ρ-effective-area correction is the classical PU subtraction used in photon isolation. PUPPI (PileUp Per Particle Identification) reweights PF candidates by a PU probability and is used broadly in CMS for jets and missing transverse energy; its application specifically to photon-related isolation quantities is not documented as a production photon technique in the sources reviewed **[unverified]**.

5.7 Anomalous-signal (spike) rejection

Production rejection is topological (Swiss-cross E_4/E_1, strip fine-grain veto bit) plus timing (Section 3.6). ML approaches to distinguish genuine showers from spikes are a natural extension but are not documented as production methods in the sources reviewed.

5.8 Classification in flagship analyses: event/analysis-level ML

$H \to \gamma\gamma$ uses a photon-ID BDT (object level) and, at the event level, a diphoton BDT for signal/background separation and category definition. $HH \to b\bar{b}\gamma\gamma$ combines object-level ML (photon ID, b-tagging) with event-level BDTs/DNNs exploiting the full $b\bar{b}\gamma\gamma$ kinematics; the ATLAS Run 2+partial-Run 3 $b\bar{b}\gamma\gamma$ analysis, for example, trained BDTs in kinematic regions and fit $m_{\gamma\gamma}$ across 14 signal regions. The key conceptual distinction: photon-reconstruction ML produces the photon four-vector, its uncertainty, and an ID score; analysis-level ML consumes these as inputs alongside other reconstructed objects.

5.9 Summary table

6 Foundation Models for Photon Physics: Current State

Table 1: Machine-learning techniques across the CMS photon reconstruction and analysis chain.

Task	**Traditional method**	**ML method (current/emerging)**	**Performance gain reported**	**Ref.**
Superclustering	Geometric Mustache window	DeepSC graph NN + self-attention	Improved energy resolution & PU/noise robustness, especially $1.0 < \|\eta\| < 1.5$	[22]
Photon energy regression	Analytic containment corrections	Semi-parametric BDT / DNN regression with per-photon σ_E	Sub-percent barrel resolution; per-event uncertainty	[1]
Photon ID / π^0 rejection	Cut-based shower-shape + isolation	BDT on high-level vars; CNN on ECAL crystal images	CNN outperforms kinematic-only; resolves merged photons	[3,23,24]
Diphoton vertex selection	Track-recoil heuristics	Vertex-ID + vertex-probability BDTs	∼75–80% within-1 cm efficiency	[7,15]
Per-event mass resolution	Global resolution model	Diphoton σ_m/m BDT	Sharper categorization, improved sensitivity	[16,10]
Spike rejection	Swiss-cross E_4/E_1 + timing	(ML not in production)	—	[17]
Event categorization ($H \to \gamma\gamma$, $HH \to b\bar{b}\gamma\gamma$)	Cut-based categories	Event-level BDT/DNN	Improved S/B per category	[8,9]

6.1 Reference point: jet foundation models

The HEP foundation-model literature is overwhelmingly jet-centric. Representative works include the Particle Transformer (ParT), a transformer for jet tagging using pairwise particle-interaction features, trained on the JetClass dataset of 100M jets; OmniLearn, a jet model simultaneously facilitating classification, generation, and regression, explicitly described by its authors as a foundation model, with a billion-jet upgrade (OmniLearned); OmniJet-α, the first cross-task foundation model for particle physics, using a VQ-VAE tokenizer and autoregressive generative pretraining; and Masked Particle Modeling (MPM), a masked-modeling self-supervised scheme for permutation-invariant particle sets, with a follow-up questioning whether tokenization is necessary at all. Other paradigms include resimulation-based self-supervision (RS3L), HEP-JEPA, and pretrained event classifiers. These operate on particle-cloud/jet-constituent representations, not on calorimeter images, and none is applied to photons or the ECAL specifically.

6.2 Detector-data foundation models (non-photon)

FM4NPP is a self-supervised scaling foundation model with up to 188M parameters, built on Time Projection Chamber tracking data for track finding, particle ID, and noise tagging at the sPHENIX experiment — not ECAL, not photons. It demonstrates the pretrain-then-adapt (frozen backbone with lightweight adapters) paradigm on real detector data. A self-supervised multimodal pretraining strategy for heterogeneous neutrino detectors augments masked reconstruction with relational voxel-level objectives and includes calorimetric streams, but targets a neutrino-detector concept (FASERCal), not CMS ECAL photon tasks.

6.3 Calorimeter generative and adjacent-repurposable models

A substantial body of work generates calorimeter showers for fast simulation and is architecturally repurposable as pretrained encoders, though built for generation rather than representation learning: CaloGAN, the CaloChallenge 2022 benchmark and its datasets, CaloFlow, image-based CaloDiffusion, and point-cloud CaloClouds/CaloClouds II.

Three recent works go further and demonstrate genuine pretraining-and-transfer specifically for EM/calorimeter generation:

- **CaloDiT-2**, a diffusion-transformer pretrained across multiple detector geometries and adapted to new ones with up to $25\times$ less data and $20\times$ faster training, explicitly framed by its authors as a first step toward a foundation model for fast simulation;
- **Cross-geometry transfer learning** in fast EM shower simulation, pretraining a point-cloud generative model on the ILD detector and fine-tuning to new geometries, achieving a 44% improvement in Wasserstein distance with only 100 target-domain samples using parameter-efficient bias-only fine-tuning (updating 17% of parameters);
- a **Mixture-of-Experts calorimetry foundation model** generating photon showers in multiple materials by adapting FM4NPP (a forward-looking 2026 preprint).

All three are generative fast-simulation models, not photon-ID or representation-learning models.

6.4 Verified gap

Based on a dedicated literature search, **no published photon-specific or CMS-ECAL-specific foundation model exists** in the sense of self-supervised/masked pretraining plus fine-tuning/transfer for photon identification, energy regression, or ECAL crystal-level representation learning applied to downstream physics tasks. The only genuine SSL-pretraining-plus-transfer detector-data models identified are on tracking (FM4NPP), jets (OmniLearn family, OmniJet-α, MPM), or neutrino detectors. All ECAL/photon-specific ML with pretraining/transfer identified is generative fast simulation (CaloDiT-2, cross-geometry transfer, MoE calorimetry FM). All CMS ECAL crystal-image photon/π^0 classification work identified is supervised and trained from scratch.

> *Caveat:* the most on-point calorimeter-transfer preprints are recent, and some carry forward-looking 2026 arXiv identifiers; their quantitative claims are the authors' own and are not yet independently validated.

6.5 Task specification for a photon foundation model: an open design question

6.5.1 Input representation

Three genuine options exist: (i) crystal-level ECAL energy/timing maps (image-like, e.g. the 32×32 seed-centered matrix already used in CMS end-to-end work), preserving the finest information; (ii) supercluster-level high-level features (compact but lossy); or (iii) a hybrid point-cloud of ECAL rechits and PF candidates in a cone, which naturally accommodates conversions and preshower hits.

6.5.2 Pretraining objectives

Candidates include masked-crystal/masked-energy modeling (predicting masked crystal energies from neighbors, the calorimeter analog of masked language modeling), contrastive learning that pulls together augmented views of the same shower while separating prompt-γ from π^0-fake classes, and multi-task pretraining jointly on energy, identification, and vertex-pointing targets.

6.5.3 The photon/jet asymmetry

A photon differs fundamentally from a jet: it is a single, small, well-defined EM shower rather than a variable-multiplicity constituent collection. The "foundation model" question is therefore genuinely open — should it be a shower-level representation learner (one shower in, one embedding out), a whole-ECAL-event representation learner (all showers, spikes, and pileup jointly), or something intermediate? This report treats the question as unresolved and flags it as the primary conceptual design decision for future work.

6.6 Evaluation metrics for a photon foundation model

Distinct from single-task benchmarks, a photon foundation model should be evaluated on:

- downstream transferability from one shared frozen or fine-tuned representation across energy regression, π^0/prompt-γ rejection, conversion classification, and vertex pointing;
- background rejection (π^0 vs. prompt-γ) at fixed signal efficiency relative to current MVA/CNN baselines;
- energy-scale and resolution calibration accuracy, and its stability across pileup and radiation-damage-induced aging — a photon-specific generalization axis with no jet analog;
- label/fine-tuning data efficiency, the core foundation-model metric: how few labeled showers are needed to match a from-scratch model, as demonstrated for jets by MPM and for calorimeter generation by the transfer-learning results above;
- robustness to data/MC discrepancy (domain adaptation), given that ECAL response evolves across the run due to aging.

6.7 Downstream usage in flagship analyses

$H \to \gamma\gamma$. A shared pretrained photon representation could improve mass resolution simultaneously through better vertex identification and energy regression, improve background rejection through stronger π^0 discrimination, and reduce reliance on several separately trained BDTs by supplying a common backbone. One methodological problem must be flagged explicitly: propagating systematic uncertainties from a foundation-model-derived quantity into the final profile-likelihood fit is unsolved. The current chain has well-understood per-object scale/smearing systematics, whereas a learned representation's uncertainty and its data/MC transferability would require a new, defensible treatment before use in a precision measurement.

$HH \to b\bar{b}\gamma\gamma$. The same photon-level benefits apply, with the additional possibility of a joint b-jet/photon event-level foundation model exploiting $b\bar{b}\gamma\gamma$ final-state correlations. If any component were intended for trigger-level (HLT or L1) use, computational latency would become a hard design constraint.

7 Research Gaps and Action Items

This section is the author's original synthesis and gap analysis. Claims about what currently exists are sourced to Section 6; the prioritization, proposed research projects, and speculative methods below are the author's own reasoning and are labeled accordingly.

7.1 Landscape assessment

(a) Photon-specific foundation-model work that already exists: None (verified in Section 6). This is a genuine white space.

(b) Adjacent/repurposable work: Substantial. Jet foundation models (OmniLearn/OmniLearned, OmniJet-α, MPM, ParT) supply transferable architectures and self-supervised recipes. Generative calorimeter models (CaloClouds/II, CaloDiffusion, CaloGAN, CaloFlow) supply encoders and point-cloud/image representations of EM showers, and CaloDiT-2 together with the cross-geometry transfer work supplies concrete evidence that pretraining and fine-tuning transfer across calorimeter geometries with large data-efficiency gains. FM4NPP supplies a proof that SSL-pretrained detector-data models with adapters outperform from-scratch baselines. CMS end-to-end CNNs supply labeled ECAL crystal-image pipelines and a "domain continuation" technique for MC→data transfer.

(c) Genuinely unaddressed problems: A self-supervised representation of ECAL showers usable for

discrimination/regression (not generation); a single backbone serving energy regression, identification, vertex pointing, and conversion classification jointly; a foundation model explicitly designed and evaluated for stability against radiation-damage-driven detector aging; and calibrated, fit-ready uncertainty outputs from a learned photon representation.

7.2 Promising near-term approaches (usable with existing methods)

1. **Fine-tune a CMS end-to-end CNN backbone for π^0/γ separation.** Task: prompt-γ vs. π^0/η fake discrimination. Existing tool: the CMS crystal-image pipeline, combined with standard vision backbones. Difficulty: low. Expected value: high — directly targets the dominant background, with an existing labeled pipeline.
2. **Masked-image-modeling pretraining on ECAL crystal maps.** Task: self-supervised pretraining of a vision transformer or calorimeter CNN on unlabeled ECAL seed-crystal images, followed by fine-tuning for identification and regression. Existing tool: MAE/ViT codebases adapted to ECAL geometry. Difficulty: medium. Expected value: high — the first genuine photon representation learner; measure label efficiency against from-scratch training.
3. **Transfer a point-cloud calorimeter encoder (CaloClouds-style) to discrimination.** Task: repurpose a generative encoder's latent space for π^0/γ and conversion classification. Existing tool: CaloClouds II. Difficulty: medium. Expected value: medium-to-high — tests the "generative model as pretrained encoder" hypothesis on EM showers.
4. **Adapt a jet foundation-model tokenizer/transformer (OmniJet-α/MPM) to ECAL point clouds.** Task: treat ECAL rechits plus PF candidates in a cone as a set, pretrain with MPM, fine-tune for photon tasks. Existing tools: MPM, OmniJet-α. Difficulty: medium-to-high. Expected value: high — leverages a mature self-supervised recipe on a new data modality.
5. **Learned per-event diphoton mass-resolution/vertex model with calibrated uncertainty.** Task: replace or augment the σ_m/m and vertex-probability BDTs with a compact transformer over recoil tracks, conversion information, and per-photon embeddings. Existing tool: existing CMS vertex-BDT inputs. Difficulty: medium. Expected value: high — improves the mass fit directly.

7.3 Underexplored / not-yet-attempted directions

- A **joint photon–jet event-level foundation model** exploiting $b\bar{b}\gamma\gamma$ correlations, unifying object embeddings across subdetectors.
- **Physics-informed pretraining objectives** specific to EM shower development — for example, predicting the longitudinal shower profile or shower-maximum depth as an auxiliary self-supervised target, directly encoding the physics of pair-production cascades.
- **Multimodal pretraining** combining ECAL crystal images with tracker conversion information, so the model natively handles the dual reconstruction path of converted photons.
- **Uncertainty-aware, calibrated foundation-model outputs** designed for direct use in profile-likelihood fits, addressing the systematics-propagation problem flagged in Section 6.7.
- **Aging-robust foundation models** via continual or online pretraining across data-taking eras, explicitly trained to be invariant to transparency loss and noise growth — the photon-specific generalization axis identified above.

7.4 Concrete action items for a ML4HEP researcher

1. **π^0/γ crystal-image classifier benchmark.** Build a public benchmark and baseline for prompt-γ vs. π^0 discrimination from ECAL images. Inputs: CMS Open Data (2016 NanoAOD plus AOD/MiniAOD

for crystal-level rechits) or simulation; modest GPU budget. Deliverable: benchmark dataset and baseline paper.

2. **ECAL masked-image-modeling backbone.** Pretrain a ViT/CNN with masked autoencoding on unlabeled ECAL seed-crystal images; publish the backbone and label-efficiency curves for identification and regression. Inputs: large unlabeled shower sample; multi-GPU. Deliverable: pretrained model and public tool.
3. **Multi-task photon head.** Fine-tune the backbone jointly for energy regression, identification, and vertex pointing; quantify whether shared pretraining outperforms per-task BDTs. Inputs: labeled MC. Deliverable: paper plus CMSSW-compatible ONNX model.
4. **Generative-encoder transfer study.** Quantify how well CaloClouds/CaloDiffusion latent representations transfer to discrimination tasks. Inputs: CaloChallenge datasets plus ILD/CMS showers. Deliverable: comparison paper.
5. **Calibrated uncertainty for the diphoton fit.** Produce and validate a learned σ_m/m estimator with fit-ready uncertainties; demonstrate a systematic-propagation treatment. Inputs: $H \to \gamma\gamma$ MC plus $Z \to ee$ data. Deliverable: methods paper.
6. **Aging-robustness evaluation framework.** Define and release a protocol testing photon-model stability across simulated transparency loss and noise growth. Inputs: aged-detector MC campaigns. Deliverable: benchmark plus paper.
7. **Conversion-classification ML.** Train a dedicated classifier for converted vs. unconverted photons and conversion recovery. Inputs: MC with conversion truth. Deliverable: tool plus technical note.
8. **Trigger-latency feasibility study.** Assess whether a compact, distilled photon backbone can run within HLT/L1 latency budgets. Inputs: timing benchmarks on target hardware. Deliverable: feasibility report.

7.5 New methods that do not yet exist (speculative, foundation-model-specific)

- A **shower-level masked-crystal transformer** ("Calo-MAE for photons"), whose pretraining objective is reconstructing masked crystal energies and timing, producing a photon embedding transferable to all downstream photon tasks — the missing direct analog of MPM for single EM showers.
- A **physics-conditioned contrastive objective** treating Geant4 resimulations of the same incident photon (varying only stochastic shower development) as positive pairs, an EM-shower analog of RS3L, yielding representations invariant to shower fluctuations but sensitive to particle identity.
- An **aging-equivariant architecture** with an explicit conditioning input for detector-response state (laser-transparency correction), trained so embeddings are invariant to aging by construction.
- A **dual-path converted-photon foundation model** that jointly ingests ECAL image and conversion-track streams with cross-attention, natively resolving the double-counting/efficiency tradeoff of converted photons.

8 Datasets

8.1 Existing datasets usable today

- **CMS Open Data (Run 2 legacy, 2016).** Primary datasets in AOD/MiniAOD and, from 2016 onward, NanoAOD, hosted on the CERN Open Data Portal. NanoAOD stores photon objects as flat ROOT branches (pre-computed identification/energy, limited precision, roughly $20\times$ smaller than MiniAOD; PF candidates dropped except in special `nanoaod-pf` derived samples). Crystal-level ECAL rechits — required for image-based photon work — are available only in AOD/MiniAOD,

analyzed via CMSSW. Suitability: fine-tuning and evaluation; the 2012 Open Data was already used for end-to-end $H \to \gamma\gamma$ CNNs.

- **Centrally produced CMS MC (internal).** γ+jet, diphoton (SM continuum), $H \to \gamma\gamma$, and $HH \to b\bar{b}\gamma\gamma$ samples exist in CMS central production, accessible internally via DAS/McM, generally not public. Suitability: fine-tuning/evaluation for CMS members; not a public pretraining corpus.
- **CaloChallenge 2022 datasets.** Public Geant4 shower datasets of increasing dimensionality. Dataset 1 (ATLAS-derived) contains photon and charged-pion showers ($\approx$ 121,000 photon showers, 368 voxels over 5 layers, incident energies 256 MeV–4.2 TeV); Datasets 2 and 3 contain electron showers (100k each, 1–1000 GeV) at higher granularity. Suitability: pretraining and benchmarking of calorimeter representations, though not CMS-ECAL geometry and not π^0-vs-γ labeled.
- **CaloClouds ILD photon showers.** Point-cloud photon showers (up to $\sim$6000 points, 10–90 GeV) for the ILD ECAL, with public code; the public ILD set has $\approx$ 23,413 photon showers at 100–1000 GeV, extendable. Suitability: pretraining/architecture prototyping; not CMS geometry.
- **Public photon-ID/regression benchmarks.** No widely adopted CMS photon-ID/regression benchmark dataset exists publicly; the closest analogs are the phenomenology calorimeter datasets above and the CMS Open Data pipelines. A dedicated public π^0/γ benchmark is a gap (Action Item 1 above).

Table 2: Dataset inventory for photon foundation-model research.

Dataset	Contents	Scale	Access	Suitability
CMS Open Data 2016 (AOD/MiniAOD)	Photon objects + crystal-level ECAL rechits, PF	Multi-fb^{-1} collision data	opendata.cern.ch (CMSSW)	Fine-tuning, evaluation, crystal-image ML
CMS Open Data 2016 (NanoAOD)	Flat photon branches (ID/energy)	Same events, flat trees	opendata.cern.ch (ROOT/Python)	Fine-tuning/eval on high-level features
CMS central MC (γ+jet, $\gamma\gamma$, $H \to \gamma\gamma$, $HH \to b\bar{b}\gamma\gamma$)	Full-sim signal/background	Large (internal)	DAS/McM (CMS internal)	Fine-tuning/eval (internal)
CaloChallenge 2022 DS1 (photons)	Geant4 photon showers, 368 voxels	$\sim$121,000 showers	Zenodo 10.5281/zenodo.8099322	Pretraining/benchmark (non-CMS)
CaloChallenge DS2/DS3 (electrons)	Higher-granularity EM showers	100k each	Zenodo 10.5281/zenodo.6366271	Pretraining/benchmark (non-CMS)
CaloClouds ILD photons	Point-cloud photon showers	$\sim$23,413 (public)	GitHub (FLC-QU-hep)	Architecture prototyping (non-CMS)
Public CMS π^0/γ ID benchmark	—	—	Does not exist	Gap (to be created)

8.2 Datasets that would need to be generated

8.2.1 Required physics processes

- **Inclusive γ+jet across a broad p_T spectrum** — the workhorse for prompt-γ vs. π^0-fake pretraining, enriching both classes simultaneously.
- **$H \to \gamma\gamma$ and $HH \to b\bar{b}\gamma\gamma$ signals at multiple mass/coupling points** — for downstream fine-tuning, with multiple κ_λ points for the HH case.

- $Z \to \ell\ell\gamma$ **(radiative** Z**)** — for calibration-style pretraining and validation with a known photon.
- $\pi^0/\eta \to \gamma\gamma$**-enriched QCD (dijet) samples** — for hard-negative mining of the merged-photon fake.

8.2.2 Statistics

Self-supervised pretraining at foundation-model scale plausibly requires $\mathcal{O}(10^7$–$10^8)$ shower objects to approach the label-free scaling seen in jet foundation models (JetClass comprises 100M jets; OmniLearned was trained on more than 1B jets). Fine-tuning and evaluation require far less, on the order of 10^5–10^6 labeled showers per task, consistent with the transfer-learning data-efficiency results reported for calorimeter transfer (44% gains with only 100 target-domain samples; $25\times$ data reduction). This estimate is the author's own order-of-magnitude reasoning, tied to the purpose of each sample.

8.2.3 Toolchain and simulation strategy

Hard-process generation: MadGraph5_aMC@NLO or POWHEG. Parton shower and hadronization: PYTHIA8. Detector simulation is the pivotal choice: crystal-level pretraining and π^0/γ image work require full Geant4-based CMS detector simulation (CMSSW) to capture per-crystal shower detail, followed by digitization and reconstruction; fast simulation (CMS FastSim or Delphes) is inadequate for crystal-level shower shapes but tolerable for high-level-feature or kinematic tasks and for large-statistics pretraining of coarse representations.

8.2.4 Action plan

1. Generate the hard process.
2. Shower and hadronize.
3. Run full Geant4 CMSSW detector simulation for crystal detail (or FastSim for coarse/kinematic tasks).
4. Digitize and reconstruct to rechits, superclusters, and PF candidates.
5. Truth-match and label (prompt-γ vs. π^0/η, conversion flags, true vertex).
6. Export to an ML-friendly format (crystal-image tensors and/or point clouds).

Computational note. Full Geant4 simulation at $\mathcal{O}(10^7$–$10^8)$ showers is computationally expensive. The pragmatic route is to repurpose and augment existing centrally produced CMS γ+jet/diphoton samples for pretraining, reserving fresh full-simulation production for labeled fine-tuning/evaluation sets and for dedicated aging-scenario studies. The CaloDiT-2 and cross-geometry transfer results (Section 6.3) suggest that generative augmentation could further reduce the required full-simulation statistics.

References

Section 1 — Photon Basics, Detector, and ECAL

[1] CMS Collaboration, "Electron and photon reconstruction and identification with the CMS experiment at the CERN LHC," *JINST* **16** (2021) P05014, doi:10.1088/1748-0221/16/05/P05014, arXiv:2012.06888.

[2] CMS Collaboration, "Performance of the CMS electromagnetic calorimeter in pp collisions at $\sqrt{s} =$ 13 TeV," *JINST* **19** (2024) P09004, doi:10.1088/1748-0221/19/09/P09004, arXiv:2403.15518.

[3] CMS Collaboration, "Performance of electron reconstruction and selection with the CMS detector in pp collisions at $\sqrt{s} = 8$ TeV," *JINST* **10** (2015) P06005, arXiv:1502.02701.

[4] CMS ECAL Group, "Radiation hardness qualification of $PbWO_4$ scintillation crystals for the CMS ECAL," arXiv:0912.4300.

[5] CMS ECAL Group, "The CMS electromagnetic calorimeter: scintillation properties," arXiv:0810.0381.

[6] CMS Collaboration, "The CMS experiment at the CERN LHC," *JINST* **3** (2008) S08004.

Section 2 — Physics Motivation

[7] CMS Collaboration, "Observation of the diphoton decay of the Higgs boson and measurement of its properties," *Eur. Phys. J. C* **74** (2014) 3076, arXiv:1407.0558.

[8] CMS Collaboration, "Measurements of Higgs boson production cross sections and couplings in the diphoton decay channel at $\sqrt{s} = 13$ TeV," arXiv:2103.06956.

[9] CMS Collaboration, "Measurements of inclusive and differential Higgs boson production cross sections at $\sqrt{s} = 13.6$ TeV in the $H \to \gamma\gamma$ decay channel," arXiv:2504.17755.

[10] CMS Collaboration, "Search for nonresonant Higgs boson pair production in final states with two bottom quarks and two photons at $\sqrt{s} = 13$ TeV," *JHEP* **03** (2021) 257, arXiv:2011.12373.

[11] CMS Collaboration, "Search for resonant production of high-mass photon pairs," *Phys. Rev. Lett.* **117** (2016) 051802, arXiv:1606.04093.

[12] CMS Collaboration, "Measurement of the $Z\gamma$ production cross section," arXiv:2601.14102.

Section 3 — Reconstruction Challenges

[13] CMS Collaboration, "Performance of photon reconstruction and identification with the CMS detector in pp collisions at $\sqrt{s} = 8$ TeV," *JINST* **10** (2015) P08010, arXiv:1502.02702.

[14] CMS Collaboration, "The CMS trigger system," *JINST* **12** (2017) P01020, arXiv:1609.02366.

[15] CMS Collaboration, "Triggering on electrons and photons with CMS," arXiv:1202.0594.

[16] CMS Collaboration, "Search for Higgs boson pair production in the four b quark final state in proton-proton collisions," vertex-BDT documentation, arXiv:2208.12279.

Section 4 — State-of-the-Art Reconstruction

[17] B. Marzocchi et al., "Deep learning techniques for energy clustering in the CMS ECAL," arXiv:2204.10277.

[18] CMS Collaboration, "Measurements of Higgs boson production and decay rates and coupling strengths using per-event resolution," arXiv:1708.09215.

Section 5 — Machine Learning

[19] M. Andrews et al., "End-to-End Physics Event Classification with CMS Open Data: Applying Image-Based Deep Learning to Detector Data for the Direct Classification of Collision Events at the LHC," *Comput. Softw. Big Sci.* **4** (2020) 6, arXiv:1807.11916.

[20] CMS Collaboration, "Reconstruction of decays to merged photons using end-to-end deep learning with domain continuation in the CMS detector," arXiv:2204.12313.

[21] ATLAS Collaboration, "Search for Higgs boson pair production in the two bottom quarks plus two photons final state," arXiv:2507.03495.

Section 6 — Foundation Models

[22] H. Qu, C. Li, S. Qian, "Particle Transformer for Jet Tagging," arXiv:2202.03772; JetClass dataset, Zenodo 10.5281/zenodo.6619768.

[23] V. Mikuni, B. Nachman, "OmniLearn: solving key challenges in collider physics with foundation models," *Phys. Rev. D* **111** (2025) L051504, arXiv:2404.16091.

[24] W. Bhimji et al., "OmniLearned: a foundation model framework for all tasks involving jet physics," arXiv:2510.24066.

[25] J. Birk, A. Hallin, G. Kasieczka, "OmniJet-α: the first cross-task foundation model for particle physics," arXiv:2403.05618.

[26] T. Golling et al., "Masked Particle Modeling on Sets: Towards Self-Supervised High Energy Physics Foundation Models," arXiv:2401.13537; follow-up, "Is Tokenization Needed for Masked Particle Modelling?," arXiv:2409.12589.

[27] P. Harris et al., "Resimulation-based self-supervised learning for pretraining physics foundation models (RS3L)," *Phys. Rev. D* **111** (2025) 032010, arXiv:2403.07066.

[28] J. Bardhan et al., "HEP-JEPA: A foundation model for collider physics using joint embedding predictive architecture," arXiv:2502.03933.

[29] "FM4NPP: a scaling foundation model for nuclear and particle physics," arXiv:2508.14087.

[30] "Towards foundation-style models for energy-frontier heterogeneous neutrino detectors via self-supervised pre-training," arXiv:2604.07037.

[31] M. Paganini, L. de Oliveira, B. Nachman, "CaloGAN: Simulating 3D High Energy Particle Showers in Multilayer Electromagnetic Calorimeters with Generative Adversarial Networks," *Phys. Rev. D* **97** (2018) 014021, arXiv:1712.10321.

[32] C. Krause et al., "CaloChallenge 2022: a community challenge for fast calorimeter simulation," arXiv:2410.21611.

[33] "CaloFlow: Fast and Accurate Generation of Calorimeter Showers with Normalizing Flows," arXiv:2106.05285; extension arXiv:2210.14245.

[34] "CaloClouds: Fast Geometry-Independent Highly-Granular Calorimeter Simulation," arXiv:2305.04847; "CaloClouds II: ultra-fast geometry-independent highly-granular calorimeter simulation," arXiv:2309.05704.

[35] "A Generalisable Generative Model for Multi-Detector Calorimeter Simulation" (CaloDiT-2), arXiv:2509.07700.

[36] "Cross-Geometry Transfer Learning in Fast Electromagnetic Shower Simulation," arXiv:2512.00187.

[37] "Generalizable Foundation Models for Calorimetry via Mixtures-of-Experts and Parameter Efficient Fine Tuning," arXiv:2603.28804 (forward-looking preprint).

Section 8 — Datasets

[38] CERN Open Data Portal, CMS documentation, https://opendata.cern.ch/docs/about-cms.

[39] CaloChallenge 2022 dataset repository, Zenodo 10.5281/zenodo.8099322 (Dataset 1) and 10.5281/zenodo.6366271 (Datasets 2/3).

[40] CaloClouds public code and data (FLC-QU-hep), associated with arXiv:2305.04847 and arXiv:2309.05704.

*Prepared July 2, 2026. Claims about existing methods are sourced to CMS public results and the arXiv/DOI references above; Section 7 is explicitly the author's original synthesis. Items labeled **[unverified]** could not be confirmed against a public primary source at the time of writing.*