RADIOLOGY · WHITEPAPER

One Encoder for Every CT: The HOPPR CT Foundation Model

One 3D CT model for chest, abdomen/pelvis and head CT, benchmarked against eight open-weight models

Dawei Gui, Evin Tunador, Akash Pattnaik, Ishan Mishra and Harris Bergman · HOPPR · November 2026

> *The model described here is a research model. It is not cleared or approved by the U.S. FDA and is not intended for diagnostic use. Developers who build on it must validate performance in their own product and setting.*

HOPPR has trained a single 3D CT foundation model that serves chest, abdomen/pelvis and head CT. The encoder supports variable volume dimensions and whole-volume evaluation within defined input limits. Scans do not have to be resized to one fixed cube. It learned from HOPPR's de-identified CT images and their radiology reports. No manual annotations were commissioned to train it.

We compared it with eight open-weight models on four benchmarks: seven CT foundation models and the general-purpose video model V-JEPA 2.1-L. Every model was frozen and evaluated with the same linear-probe protocol. Our model has the highest score on all four:

- **HOPPR internal benchmark (75 findings).** Mean AUROC 0.861, against 0.823 for the strongest open-weight model (SPECTRE). It is first in every region: chest, abdomen/pelvis and head.
- **CT-RATE (public chest CT).** Macro AUROC 0.874, ahead of COLIPRI (0.866), NV-Reason-CT (0.866) and SPECTRE (0.849), all three pretrained on CT-RATE. HOPPR's encoder was not directly pretrained or report-aligned on CT-RATE, though some of the models it learned from were (see "About the scores").
- **RSNA intracranial hemorrhage (public head CT).** Macro AUROC 0.930 over five hemorrhage targets, and any-hemorrhage AUROC 0.954, ahead of every other model by at least 0.023.
- **RSNA abdominal trauma (public abdominal CT).** Macro AUROC 0.840 on the primary readout. It is ahead of SPECTRE (0.831), though that margin is within the confidence interval. On the secondary readout it scores 0.867, ahead of every other model by at least 0.034.

On the first three benchmarks, and on the secondary trauma readout, the paired 95% interval for the difference excludes zero against every other model.

Pillar-0 was evaluated with its chest CT checkpoint on every benchmark, so its head and abdomen results do not represent Pillar-0's region-specific models. nnFoundationViT appears on the internal benchmark only.

One backbone that is strong in every region means every CT product we build can start from the same foundation. Each one inherits improvements to it, without retraining a vision model for every region.

What the model does

CT is not one image type. A head CT covers about 15–20 cm of anatomy at 0.5 mm in-plane resolution. A chest–abdomen–pelvis study can span 60 cm and hundreds of slices. Scans differ in slice thickness, reconstruction kernel and contrast phase. Many CT foundation models handle this variety by resampling every scan to one fixed cube, or by specialising in one body region.

The HOPPR CT foundation model is built differently:

- **Variable volume size.** The encoder accepts volumes of different depth and in-plane size. Each scan is embedded as a whole volume, up to 256 slices × 512 × 512 at HOPPR's standard voxel spacing. Appendix A gives the preprocessing and what happens to longer scans.
- **One model for every region.** A single model covers chest, abdomen/pelvis and head CT, with or without contrast.
- **Radiologist windows as input.** The model reads CT as four standard Hounsfield-unit windows: soft tissue, bone, lung and brain.
- **Learned from images and reports.** The model was first trained with self-supervised learning and knowledge distillation from existing CT models, then aligned with the corresponding radiology reports. Report-derived finding labels were also used to balance which studies were sampled during alignment. No newly commissioned manual annotations were used to train the encoder.
- **A ready-to-use embedding.** Each scan becomes one 1,024-dimensional vector that downstream models can use directly.

How we evaluated

We measure representation quality with a **frozen linear probe**, so the score reflects the features rather than the capacity of a task head. Every model is evaluated the same way:

- **One embedding per scan.** The encoder is frozen and embeds each scan once, as a whole volume, in that model's own native preprocessing (Appendix C).
- **One classifier per finding.** We fit an L2-regularised logistic regression on standardised features. The regularisation strength is chosen from 11 values by patient-grouped 5-fold cross-validation inside the training data only.
- **Scores and uncertainty.** We report AUROC averaged over findings. Confidence intervals and model-to-model differences use a patient-cluster bootstrap. For differences, every model is scored on the same resampled patients (a paired comparison).

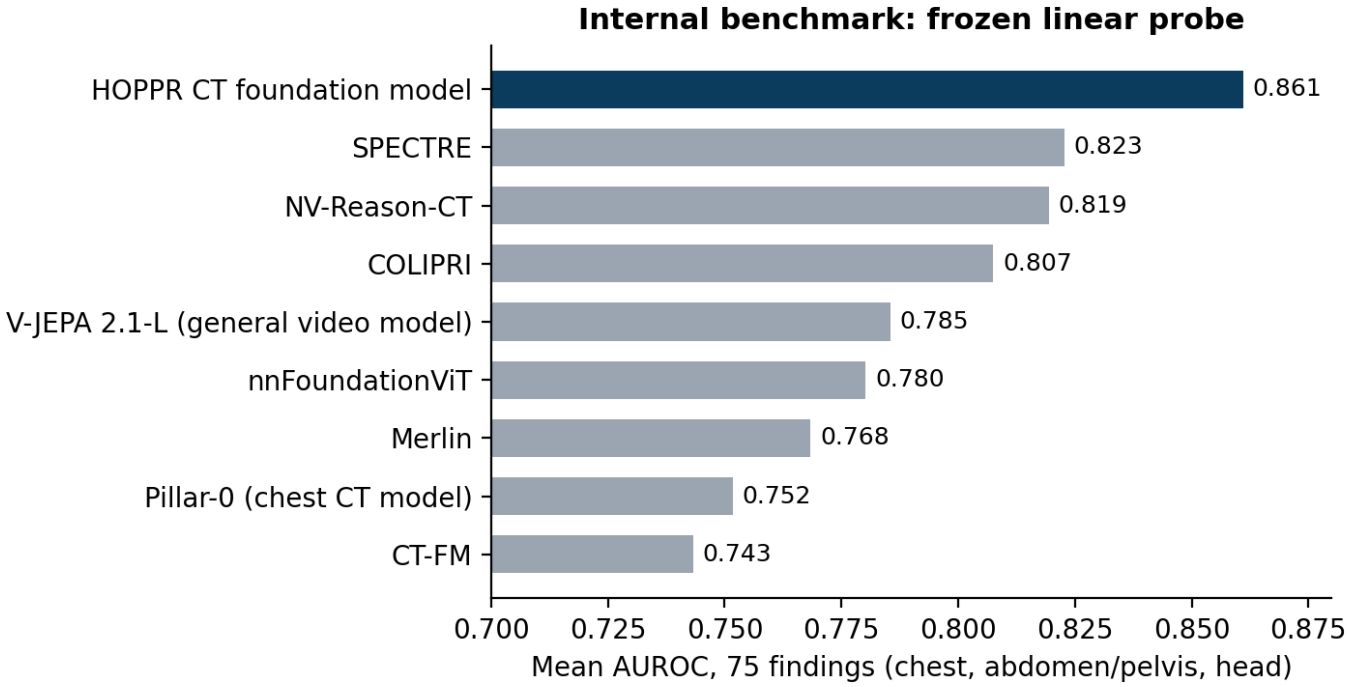
- **Same pipeline for everyone.** HOPPR computed all scores in this paper, for all models, with this pipeline.

Benchmark	Region	Targets	Probe training data	Evaluation set
HOPPR internal	chest, abdomen/pelvis, head	75 findings (9 / 57 / 9)	up to 20,000 studies per region and repeat, from a pool of HOPPR studies	development split: 57,975 chest, 23,582 abdomen/pelvis and 8,323 head studies (+1,886 for acute hemorrhage)
CT-RATE [1]	chest	18 findings	official training split: 24,128 scans from 20,000 patients	official validation split: 3,039 volumes (1,564 scans, 1,304 patients)
RSNA ICH, derived [2]	head	5 hemorrhage targets	17,009 series from 14,818 patients	2,135 series from 1,848 patients
RSNA abdominal trauma (RATIC), derived [3]	abdomen/pelvis	6 injury targets	4 of 5 folds	3,147 patients (4,711 series), 5-fold patient cross-validation

Appendix B describes each benchmark's protocol.

Results

Internal benchmark: chest, abdomen/pelvis and head CT



Mean AUROC on the HOPPR internal benchmark, frozen linear probe, development split.

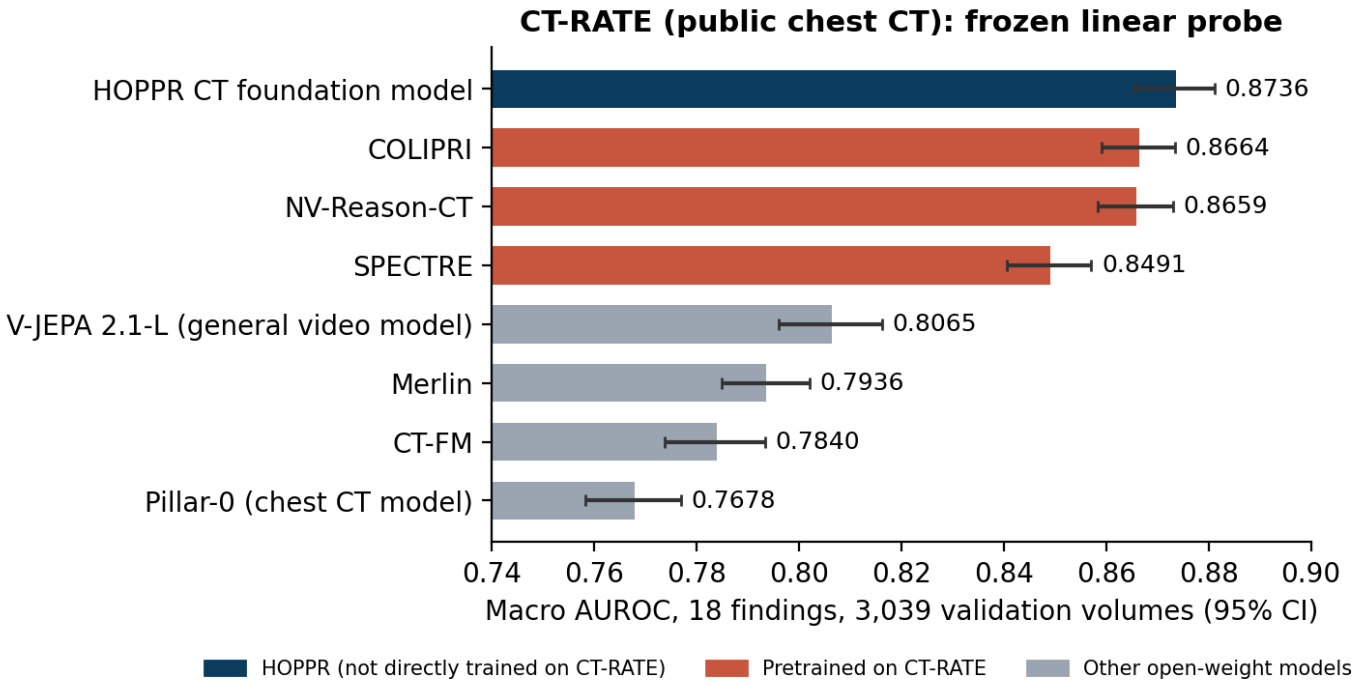
Model	Overall	Difference from ours (95% CI)	Chest (9)	Abdomen/pelvis (57)	Head (9)
HOPPR CT foundation model	**0.8611**	—	**0.8817**	**0.8486**	**0.8530**
SPECTRE [4]	0.8228	−0.0383 (−0.0401, −0.0366)	0.8479	0.8134	0.8070
NV-Reason-CT [5]	0.8194	−0.0417 (−0.0437, −0.0398)	0.8637	0.7948	0.7996
COLIPRI [6]	0.8075	−0.0536 (−0.0558, −0.0515)	0.8581	0.7792	0.7851
V-JEPA 2.1-L, general video model [7]	0.7855	−0.0756 (−0.0775, −0.0737)	0.7908	0.7513	0.8145
nnFoundationViT [8]	0.7802	−0.0809 (−0.0830, −0.0789)	0.7949	0.7495	0.7960
Merlin [9]	0.7684	−0.0928 (−0.0953, −0.0902)	0.7813	0.7781	0.7456
Pillar-0, chest CT model [10]	0.7518	−0.1093 (−0.1117, −0.1069)	0.7522	0.7298	0.7733
CT-FM [11]	0.7433	−0.1178 (−0.1203, −0.1152)	0.7588	0.7219	0.7493

Mean AUROC on the development split. Overall is the unweighted mean of the three regions. Differences are paired, with 10,000 patient-cluster bootstrap resamples.

- **A wide overall lead.** Our model leads by 0.038 overall. The paired interval excludes zero against every other model, overall and in each region separately.
- **First in every region.** Each external model is strongest in one place: NV-Reason-CT on chest, SPECTRE on abdomen/pelvis and V-JEPA 2.1-L on head. Our model leads each of them there: +0.018 on chest, +0.035 on abdomen/pelvis and +0.039 on head.
- **Abdomen/pelvis.** This 57-finding region is the lowest-scoring region for most public models. Only our model scores above 0.84 there.

These are development-split results, and the internal labels come from the same kind of reports the model was aligned with. “About the scores” explains what that means for the size of the lead.

CT-RATE: public chest CT



Macro AUROC on CT-RATE validation with 95% confidence intervals.

Model	Pretrained on CT-RATE?	Macro AUROC (95% CI)	Difference from ours (95% CI)
HOPPR CT foundation model	**not directly**	**0.8736 (0.8656–0.8811)**	—
COLIPRI	yes	0.8664 (0.8590–0.8734)	−0.0072 (−0.0105, −0.0038)
NV-Reason-CT	yes	0.8659 (0.8583–0.8730)	−0.0077 (−0.0112, −0.0041)
SPECTRE	yes	0.8491 (0.8406–0.8571)	−0.0244 (−0.0285, −0.0204)
V-JEPA 2.1-L, general video model	no	0.8065 (0.7960–0.8162)	−0.0670 (−0.0734, −0.0611)
Merlin	no	0.7936 (0.7848–0.8021)	−0.0800 (−0.0877, −0.0725)
CT-FM	no	0.7840 (0.7738–0.7934)	−0.0896 (−0.0969, −0.0826)
Pillar-0, chest CT model	no	0.7678 (0.7583–0.7770)	−0.1058 (−0.1136, −0.0983)

All 18 findings, scored on the 3,039 validation volumes. Probes are trained on CT-RATE's official training split. Differences are paired, with 10,000 patient-cluster bootstrap resamples.

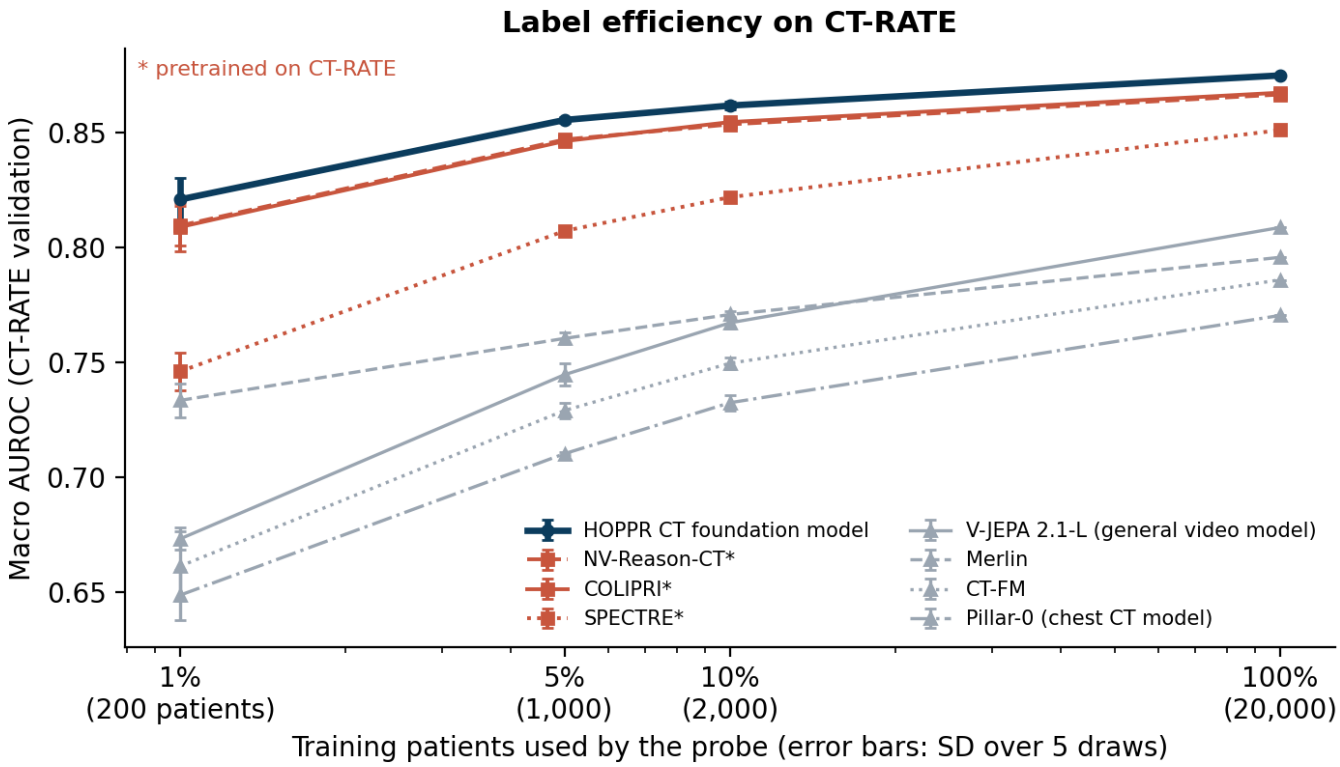
CT-RATE is the standard public benchmark for 3D chest CT. Three of the comparators were pretrained on its training split, which gives them an in-distribution advantage.

- **Ahead of the CT-RATE-pretrained models.** Our model is first, 0.007–0.008 ahead of COLIPRI and NV-Reason-CT and 0.024 ahead of SPECTRE. The individual intervals of the top three overlap. The paired

comparison, which scores every model on the same resampled patients, is what separates them, and each of its intervals excludes zero.

- **Ahead of the other models.** Among models not pretrained on CT-RATE, it leads by 0.067 or more.
- **Indirect exposure.** HOPPR's encoder was not directly pretrained or report-aligned on CT-RATE. But some of the models it learned from during distillation, SPECTRE and COLIPRI, were pretrained on CT-RATE.

Label efficiency. We repeated the CT-RATE probe with 1%, 5% and 10% of the training patients. Each size uses 5 random draws, the draws are nested (each 1% draw is inside the matching 5% draw, and so on), and every model uses the same draws.



Macro AUROC on CT-RATE as the number of training patients shrinks from 100% to 1%.

Model	1% (200 patients, ≈240 scans)	5% (1,000 patients, ≈1,205 scans)	10% (2,000 patients, ≈2,407 scans)	100% (20,000 patients, 24,128 scans)
HOPPR CT foundation model	**0.821 ± 0.009**	**0.856 ± 0.001**	**0.862 ± 0.002**	**0.875**
COLIPRI	0.809 ± 0.011	0.846 ± 0.001	0.855 ± 0.002	0.868
NV-Reason-CT	0.810 ± 0.009	0.847 ± 0.002	0.854 ± 0.001	0.867
SPECTRE	0.746 ± 0.008	0.807 ± 0.001	0.822 ± 0.002	0.851
V-JEPA 2.1-L	0.673 ± 0.005	0.745 ± 0.005	0.767 ± 0.003	0.809
Merlin	0.734 ± 0.008	0.761 ± 0.003	0.771 ± 0.002	0.796
CT-FM	0.661 ± 0.015	0.729 ± 0.004	0.750 ± 0.003	0.786
Pillar-0, chest CT model	0.649 ± 0.011	0.710 ± 0.001	0.733 ± 0.003	0.771

Macro AUROC on the 3,039 validation volumes; mean ± SD over the 5 draws.

The 100% column is not the same experiment as the main table above, which is why our model shows 0.875 here and 0.874 there:

- **Main table:** the protocol's standard estimate, averaged over 5 fits, each on a random 80% of training patients (about 19,300 scans).
- **100% point here:** a single fit on all 24,128 training scans. That is 25% more data, so every model scores 0.0003–0.0027 higher than in the main table.

The rankings agree between the two.

- **First at every size.** Our model has the highest score at every training-set size. In paired comparisons on the same draws, every other model is below it at every size, with the interval excluding zero.
- **Small budgets compared with full ones.** With 5% of the training patients our model scores 0.856, above SPECTRE trained on 100% (0.851). With 1% it scores 0.821, above every model not pretrained on CT-RATE even at 100% (best: V-JEPA 2.1-L, 0.809).
- **The cost of fewer labels.** Going from 100% to 1% of the training patients lowers our model's AUROC from 0.875 to 0.821, a drop of 0.054. The drop is 0.105 for SPECTRE and 0.136 for V-JEPA 2.1-L.

RSNA intracranial hemorrhage: a five-target evaluation (head CT)

Model	Macro AUROC, 5 targets (95% CI)	Difference from ours (95% CI)	Any hemorrhage
HOPPR CT foundation model	**0.930 (0.922–0.938)**	—	**0.954**
V-JEPA 2.1-L, general video model	0.906 (0.896–0.915)	−0.025 (−0.031, −0.018)	0.931
SPECTRE	0.898 (0.887–0.908)	−0.032 (−0.039, −0.026)	0.922
NV-Reason-CT	0.880 (0.869–0.892)	−0.050 (−0.058, −0.042)	0.907
COLIPRI	0.854 (0.841–0.868)	−0.076 (−0.087, −0.066)	0.877
CT-FM	0.849 (0.837–0.863)	−0.081 (−0.092, −0.071)	0.878
Merlin	0.836 (0.822–0.850)	−0.094 (−0.107, −0.083)	0.864
Pillar-0, chest CT model	0.817 (0.800–0.833)	−0.113 (−0.128, −0.100)	0.835

Primary (patch-average) readout. Differences are paired, with 2,000 patient bootstrap resamples.

This evaluation is derived from RSNA ICH 2019, not the challenge's own endpoint.

- **Targets.** The five targets are any hemorrhage and four subtypes: intraparenchymal, intraventricular, subarachnoid and subdural.
- **Epidural hemorrhage excluded.** RSNA ICH also labels epidural hemorrhage. We left it out of the macro score because the test split has only 36 epidural-positive series. Including epidural, our model's six-target macro AUROC is 0.925. On the epidural target alone it reaches 0.911 (95% CI 0.843–0.957) with the global-token readout.
- **Split.** The split is our own patient-level 80/10/10 split of the labelled challenge data. Appendix B gives the details.

Our model is first on both the macro score and any hemorrhage, and every paired interval excludes zero.

RSNA abdominal trauma: a six-target evaluation (abdomen/pelvis CT)

Model	Primary readout: patch average (95% CI)	Difference from ours (95% CI)	Secondary readout: global token (95% CI)	Difference from ours (95% CI)
HOPPR CT foundation model	**0.840 (0.827–0.854)**	—	**0.867 (0.855–0.879)**	—
SPECTRE	0.831 (0.816–0.845)	−0.009 (−0.023, +0.004)	0.833 (0.819–0.846)	−0.034 (−0.047, −0.021)
NV-Reason-CT	0.797 (0.781–0.811)	−0.043 (−0.057, −0.031)	0.790 (0.774–0.806)	−0.077 (−0.092, −0.062)
Merlin	0.779 (0.763–0.793)	−0.061 (−0.077, −0.047)	0.773 (0.756–0.787)	−0.095 (−0.111, −0.080)
V-JEPA 2.1-L, general video model	0.746 (0.729–0.762)	−0.094 (−0.111, −0.079)	0.746 (0.729–0.762)	−0.121 (−0.139, −0.105)
COLIPRI	0.745 (0.728–0.760)	−0.096 (−0.111, −0.081)	0.740 (0.723–0.758)	−0.127 (−0.143, −0.111)
CT-FM	0.701 (0.684–0.719)	−0.139 (−0.156, −0.121)	0.656 (0.637–0.674)	−0.212 (−0.232, −0.191)
Pillar-0, chest CT model	0.671 (0.654–0.689)	−0.169 (−0.189, −0.151)	0.667 (0.649–0.685)	−0.201 (−0.223, −0.178)

Patient-level macro AUROC over 6 binary targets, from out-of-fold predictions on all 3,147 patients. Differences are paired, with 2,000 patient bootstrap resamples. Appendix C defines each model's global-token readout; for V-JEPA 2.1-L it is the same as the patch average.

This is our own six-target binary evaluation on the labelled RSNA 2023 Abdominal Trauma (RATIC) cohort, with patient-level 5-fold cross-validation. It is not the competition's official metric.

- **Primary readout.** Our model has the highest score, but its 0.009 lead over SPECTRE is within the confidence interval, so the two are not separated. It is ahead of every other model with intervals excluding zero.
- **Secondary readout.** Our model is first, ahead of every other model with every paired interval excluding zero. Per organ it reaches AUROC 0.924 for kidney, 0.887 for spleen and 0.873 for liver injury.

Our model scores 0.027 higher with the global token than with the patch average. Most external models differ by at most 0.01 between the two readouts. We report this as an observed difference; this benchmark does not establish its cause.

Fine-tuning

A frozen probe is a deliberately conservative test. We also fine-tuned the model lightly, training only a small fraction of its parameters together with a task head:

- **Internal benchmark (development split):** mean AUROC rose from 0.861 to 0.873. That is +0.016 on abdomen/pelvis, +0.018 on head and +0.003 on chest, with every paired 95% interval above zero.
- **CT-RATE:** it rose from 0.874 to 0.880 (paired gain +0.006, 95% CI +0.004 to +0.009). The checkpoint was chosen on a hold-out of CT-RATE training patients, and the validation split was scored once.

Fine-tuned versions of the comparators were not evaluated.

About the scores

- **Same protocol, our runs.** All numbers in this paper were computed by HOPPR with one protocol, at each model's own preprocessing. They can differ from figures in the original publications, which use other splits, heads and preprocessing.
- **What the intervals cover.** Confidence intervals come from resampling evaluation patients. They do not capture:
  - variation from retraining an encoder with a different random seed (each model here is a single trained checkpoint);
  - every effect of choosing that checkpoint.
- **How "ahead" is tested.** When we say a model is ahead with the interval excluding zero, we mean a paired comparison of two models at the 95% level, without adjustment for multiple comparisons. The smallest such margin in this paper, against COLIPRI on CT-RATE, is about four standard errors from zero. We do not treat small differences as ties or as wins by any fixed threshold. Each comparison is reported with its own interval.
- **Report-derived labels.** The internal labels are extracted from radiology reports by a language model. Our model was aligned with radiology reports from HOPPR's training data, and the studies used to fit the internal probes are among them. Label noise of this kind can affect models differently, and it can favour a model aligned with reports of the same kind. So the internal margin may overstate our model's advantage. The internal evaluation studies themselves were never used in training.
- **Internal results are development results.** Internal scores are on the development split. The model's checkpoint was chosen on a subset of it, and on the development patients outside that subset it scores 0.861 as well. A one-time readout of the held-out internal test split has not yet been made for this model.
- **Exposure to the public datasets.** HOPPR's encoder was trained only on HOPPR data. During distillation, though, it learned from models pretrained on other data:
  - SPECTRE and COLIPRI, both pretrained on CT-RATE;
  - V-JEPA 2.1, trained on video;

    - nnFoundationViT, trained on public and institutional imaging collections;
    - an earlier HOPPR model.

  Among the comparators, COLIPRI, NV-Reason-CT and SPECTRE were pretrained on CT-RATE. SPECTRE's published data list includes neither RSNA dataset. We have not audited every comparator's, or every distillation source's, pretraining data at the checkpoint level for RSNA ICH or RATIC images. This applies in particular to the large multi-source collections behind CT-FM and nnFoundationViT.

- **Comparator scope.**
    - V-JEPA 2.1-L is a general-purpose video model, not a CT model.
    - Pillar-0 was run with its chest CT checkpoint on every benchmark. Pillar-0 also publishes head and abdomen models, so our head and abdomen results do not represent those.
    - nnFoundationViT was evaluated on the internal benchmark only. It was left out of the public-benchmark runs when their scope was fixed, so the public tables have seven comparators.
- **What a probe does and does not show.** A linear probe measures how accessible findings are in frozen features. It does not measure report quality, fine-tuned performance on a specific clinical task, or clinical utility.

What comes next

Three directions build on this foundation:

- **CT vision-language models.** The model is the vision backbone of HOPPR's CT report-generation and reasoning work, where encoder quality sets the ceiling for everything downstream.
- **Downstream models.** Classification and localisation models for chest, abdomen/pelvis and head CT now start from one shared encoder instead of one per region.
- **Held-out internal test readout.** A one-time readout of the held-out internal test split, to confirm the development results.

Conclusion

A single 3D encoder, trained on HOPPR's CT images and reports, can match or beat region-specific models across the body. It has the highest score on all four benchmarks in this paper:

- on our 75-finding internal benchmark, in every region;
- on public chest CT, ahead of three models pretrained on that dataset, and at every training-set size;
- on public head CT, ahead of every model;
- on public abdominal trauma CT, where its lead on the primary readout is within the confidence interval.

HOPPR aims to build radiology AI on foundations that are measured openly and compared fairly.

Appendix A: Model and preprocessing

- **Model.** HOPPR CT foundation model, October 2026 release (checkpoint SHA-256 `748bc7135fb6…`): a 3D vision transformer with about 300M parameters that outputs a 1,024-dimensional embedding per scan.
- **Development.** The model was trained in three steps:
  - self-supervised learning on HOPPR CT;
  - knowledge distillation from existing CT and imaging models, run during that same stage;
  - alignment with radiology reports.
- **Training data** (unique items, not training draws), all from HOPPR's de-identified training partition, which is patient-disjoint from the internal development and test splits:
  - Self-supervised stage: about 1.11 million CT series (581,000 chest, 304,000 abdomen/pelvis, 229,000 head).
  - Report alignment: 726,072 studies with reports (881,924 series).
- **Preprocessing.** DICOM series are converted to Hounsfield-unit volumes and resampled by linear interpolation to a standard grid:
  - chest and abdomen/pelvis: 0.7 × 0.7 × 2.5 mm, 512 × 512 in-plane;
  - head: 0.5 × 0.5 × 2.5 mm, 512 × 512 in-plane, with gantry tilt corrected.

  The public datasets are processed with the same pipelines: CT-RATE as chest, RATIC as abdomen/pelvis, RSNA ICH as head.
- **Input to the encoder.** The four windows are soft tissue (W400/L40), bone (W1800/L400), lung (W1500/L−600) and brain (W80/L40), each scaled to [0, 1]. The whole volume is used, with no crop.
  - Depth is resampled up to the next multiple of 16 slices, with a maximum of 256 slices.
  - Volumes with more than 256 stored slices are resampled down to 256 along the slice axis. That is about 0.3% of CT-RATE volumes.
  - The in-plane size is the stored size, up to 512 × 512.
- **Readouts.** The primary readout is the mean of the final-layer patch tokens. The secondary (global-token) readout is the final-layer class token.

Appendix B: Benchmark protocols

- **Probe.** Every benchmark uses the same per-target classifier: standardised features and L2 logistic regression. The regularisation strength comes from 11 values (10^{-5} to 1), chosen by patient-grouped, label-stratified 5-fold cross-validation within the probe's training data.
- **HOPPR internal.** 75 findings with report-derived labels (chest 9, abdomen/pelvis 57, head 8 plus acute hemorrhage).
  - Each score is the mean over 5 repeats. Each repeat fits on a random 80% of a patient-disjoint training pool, capped at 20,000 studies per region.
  - Scoring is on the development split.
  - Intervals use 10,000 patient-cluster bootstrap resamples.
- **CT-RATE.** The official 18 multi-abnormality labels, which are extracted from reports by a text classifier.
  - Training uses one volume per scan from the official training split (24,128 scans, 20,000 patients).
  - Scoring uses all 3,039 validation volumes (1,564 scans, 1,304 patients).
  - The headline score is the mean over 5 repeats on 80% patient draws. CT-RATE has no public test split, and nothing was tuned on its validation split.
- **RSNA ICH, derived.** Labels come from the 2019 challenge's labelled (stage 2 training) set and are assigned per series: a series is positive for a target if any of its slices is.
  - The split is ours: 80/10/10 by patient, stratified across the subtypes, with the seed fixed before any model was run.
  - The probe trains on the training split (17,009 series, 14,818 patients) and is scored once on the test split (2,135 series, 1,848 patients). The validation split was not used.
  - Series that failed geometry checks were excluded before splitting: 481 of 21,744 labelled series, mostly scans tilted more than about 26°.
- **RSNA abdominal trauma (RATIC), derived.** Six binary patient-level targets from the competition's labelled training cohort: liver, spleen and kidney injury of any grade, bowel injury, active extravasation and any injury.
  - The cohort is 3,147 patients and 4,711 series. A patient's embedding is the mean of its series embeddings.
  - Evaluation uses 5-fold patient cross-validation, stratified by injury, with the seed fixed in advance.
  - Hyperparameter selection is nested: the regularisation strength is chosen by cross-validation inside the four training folds of each outer fold.
  - Scores come from pooled out-of-fold predictions. This is not the competition's weighted log-loss endpoint.

Appendix C: Comparators

Every comparator was run from its published weights, with its own preprocessing. External models get the input resolution and intensity scaling they were designed for. For every model, the primary readout averages the final-layer patch tokens, or the final spatial feature map for convolutional encoders.

Model	Weights	Preprocessing / view	Embedding	Global-token readout
SPECTRE	`cclaess/SPECTRE`, large	0.75 × 0.75 mm in-plane, 1.5 mm slices, tiled 128 × 128 × 64; HU −1000 to 1000	1,080	combiner class token
COLIPRI	`microsoft/colipri`	192^3 cube at 2 mm isotropic, centred; HU −1000 to 1000	768	class token
NV-Reason-CT (vision tower)	`nvidia/NV-Reason-CT`	192^3 cube at 2 mm isotropic, centred	864	mean of the projector output
V-JEPA 2.1-L	public V-JEPA 2.1 large	384 × 384 in-plane, HU −1000 to 1000 to three channels	1,024	none; equals the patch average
nnFoundationViT (internal benchmark only)	`MIC-DKFZ/nnFoundationViT`	volume resampled into one 192^3 tile, 8^3 patches, z-score normalised	1,056	class token
Merlin	`stanfordmimi/Merlin`	RAS, 1.5 × 1.5 × 3 mm, HU −1000 to 1000, 224 × 224 × 160	2,048	512-dimensional image–text embedding
Pillar-0, chest CT model	`YalaLab/Pillar0-ChestCT`	fixed 320 mm view	384	global token
CT-FM	`project-lighter/ct_fm_feature_extractor`	3 × 1 × 1 mm, foreground crop, HU −1024 to 2048	512	global max-pooling

References

1. I. E. Hamamci et al., “Generalist Foundation Models from a Multimodal Dataset for 3D Computed Tomography” (CT-RATE), *arXiv:2403.17834*, 2024. https://arxiv.org/abs/2403.17834
2. A. E. Flanders et al., “Construction of a Machine Learning Dataset through Collaboration: The RSNA 2019 Brain CT Hemorrhage Challenge,” *Radiology: Artificial Intelligence*, 2(3):e190211, 2020. https://doi.org/10.1148/ryai.2020190211

3. J. D. Rudie et al., “The RSNA Abdominal Traumatic Injury CT (RATIC) Dataset,” *Radiology: Artificial Intelligence*, 6(6):e240101, 2024. https://doi.org/10.1148/ryai.240101

4. C. Claessens et al., “Scaling Self-Supervised and Cross-Modal Pretraining for Volumetric CT Transformers” (SPECTRE), *arXiv:2511.17209*, 2025. https://arxiv.org/abs/2511.17209

5. A. Myronenko et al., “NV-Reason-CT: 3D Visual Language Model for CT Analysis,” *arXiv:2609.27511*, 2026. https://arxiv.org/abs/2609.27511

6. T. Wald et al., “Comprehensive Language–Image Pre-training for 3D Medical Image Understanding” (COLIPRI), *ECCV*, 2026. https://doi.org/10.1007/978-3-032-37624-4_24

7. L. Mur-Labadia et al., “V-JEPA 2.1: Unlocking Dense Features in Video Self-Supervised Learning,” *arXiv: 2603.14482*, 2026. https://arxiv.org/abs/2603.14482

8. C. Ulrich Harsy et al., “nnFoundation: 3D Foundation Models for Radiology,” *arXiv:2609.26924*, 2026. https://arxiv.org/abs/2609.26924

9. L. Blankemeier et al., “Merlin: A Computed Tomography Vision-Language Foundation Model and Dataset,” *Nature*, 2026. https://arxiv.org/abs/2406.06512

10. K. K. Agrawal et al., “Pillar-0: A New Frontier for Radiology Foundation Models,” *arXiv:2511.17803*, 2025. https://arxiv.org/abs/2511.17803

11. S. Pai et al., “Vision Foundation Models for Computed Tomography” (CT-FM), *arXiv:2501.09001*, 2025. https://arxiv.org/abs/2501.09001

About the authors. Dawei Gui, Evin Tunador, Akash Pattnaik, Ishan Mishra and Harris Bergman worked together to produce the results in this paper. Dawei Gui is a machine learning scientist at HOPPR working on CT foundation models.