Concept-Level Vision–Language Alignment for 3D CT: Linear Probing for the CVPR 2026 3D-CT FM Challenge

Julien Khlaut[1,2,3][0009−0001−2891−7588], Charles Corbière[⋆1][0000−0001−8024−7553], Baptiste Callard[⋆1][0009−0006−7469−6540], Amaury Prat[⋆1][0009−0005−5474−929X], Leo Butsanets[1][0009−0005−2926−4135], and Corentin Dancette[1][0009−0007−9274−4910]

[1] Raidium, 27 rue du faubourg Saint-Jacques, 75014 Paris, France
[2] Department of Vascular and Oncological Interventional Radiology, Hôpital Européen Georges Pompidou, AP-HP, Paris, France
[3] Faculté de Santé, Université Paris-Cité, Paris, France
julien.khlaut@raidium.eu (corresponding author)

Abstract. This paper describes our submission to Task 1 (Linear Probing), **All-Data Track**, of the CVPR 2026 3D-CT Foundation Model Challenge, whose 17 binary targets span chest and abdominal CT under heavy class imbalance. We freeze **Jolia**, a 3D vision–language foundation model whose anatomy-anchored pretraining on 74,434 public chest and abdominal CT–report pairs yields, for every organ, a localized embedding alongside the global volume embedding. For each target we concatenate the global and a local organ-wise embedding and let a class-balanced linear probe select the relevant signal. On the 17 validation targets this reaches a mean **balanced accuracy of 0.7575** and a mean **AUROC of 0.7645** – mid-pack among the eight official challenge baselines: above the Curia 2D baseline on chest but below it on abdomen and overall, while using a much smaller frozen encoder (22 M parameters). The complete pipeline runs inside a < 4 GB Docker container at under 5 s per case, well within the challenge's 40 s budget.

Keywords: 3D CT foundation model · Linear probing · Vision–language pretraining · Embedding aggregation · Binary classification.

1 Introduction

P1 – Background. Classifying findings in 3D chest and abdominal CT is hard for reasons specific to the modality. A single volume spans dozens of organs at once, and a radiology report is correspondingly long and section-structured, typically describing findings organ by organ. The downstream targets of this challenge mix *global* findings that affect a large portion of the volume (e.g., *ascites*) with very *localized* ones that occupy a few voxels (e.g., *adrenal hyperplasia*, *kidney*

[⋆] These authors contributed equally.

stone). Finally, the 17 binary targets across three datasets are severely class-imbalanced. A successful 3D CT foundation model (FM) must therefore both encode the whole volume *and* preserve organ-level localization, and a successful adaptation protocol must turn that representation into a robust classifier under heavy imbalance.

P2 – Related work. Three families of 3D CT FMs have emerged. *Vision-only* models pretrain on images alone, through self-supervision or scaling: VoCo [13], CT-FM [11], 3DINO [14], and Curia [5]. *Vision–language* models align volume and report with a global CLIP-style objective: Merlin [3], CT-CLIP [7], and Pillar-0 [2]. *Hybrid* approaches such as SPECTRE [4] combine self-supervised and multi-modal signals. Adapting a frozen backbone to a downstream target is most commonly done by linear probing [6], but LoRA [9] and adapters [8] are popular lightweight alternatives. The common limitation of CLIP-style FMs is that they compress the whole report and volume into a single global token, losing the organ-level structure that localized findings depend on; vision-only FMs, conversely, lack any semantic anchoring. Methods that add spatial alignment either require external segmentation masks (fVLM [12]) or extra heuristics, and none of them systematically compare *feature-aggregation recipes* under a unified linear-probing protocol.

P3 – Our contribution. We submit **Jolia** [10], a 3D Atlas backbone (22 M parameters, 192^3-voxel input) pretrained with **ConQuer**, a *concept-level* contrastive loss $\mathcal{L}_{\text{ConQuer}}$, on 74,434 chest and abdominal CT/report pairs from CT-RATE, INSPECT, and Stanford-Abdominal-CT. The frozen model exposes both a global `[CLS]` token *and* 102 anatomical concept queries. Our contributions to this challenge are:

1. A concept-anchored 3D CT FM that, once frozen, provides 102 organ-level concept queries in addition to the global `[CLS]` token, without any segmentation supervision at pretraining time.
2. A simple linear-probing feature for these targets: we concatenate the global `[CLS]`, a cross-organ pooled query, and either the 10 task-relevant organ queries or a mask-pooled organ feature (`cls_roi`) for the 11 ROI targets that ship masks; the per-target LP head selects the relevant signal.
3. On the 17 validation targets this reaches **BalAcc 0.7575 / AUROC 0.7645**: it beats the Curia 2D baseline on chest (0.8447 vs. 0.7420) but trails it on abdomen (0.7459 vs. 0.7749) and overall (Curia 0.7711), and sits mid-pack among the eight official baselines, the gap concentrating on small organ-localized abdominal pathologies.

2 Method

Fig. 1 shows the ConQuer pretraining pipeline that produces our frozen backbone, and Fig. 2 shows the linear-probing evaluation protocol used for this challenge.

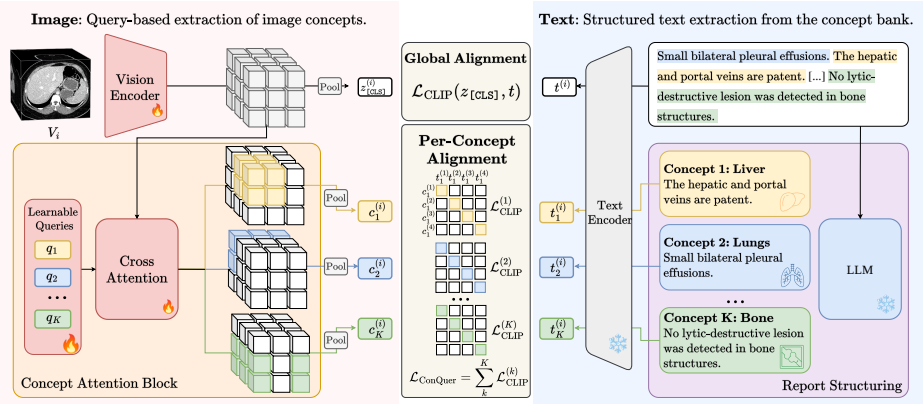

Fig. 1. Overview of ConQuer pretraining. Both the image (via learnable cross-attention queries) and the report (via concept-specific sections produced by an LLM) are decomposed into a set of concept-level representations; contrastive alignment is applied independently for each concept, in addition to the standard global CLIP loss $\mathcal{L}_{\text{CLIP}}(z_{\texttt{[CLS]}}, t)$. For this challenge the backbone is **frozen**; only the linear-probing head is trained.

2.1 Pretraining objective (Task 1, All-Data Track)

ConQuer is the vision–language pretraining method that produces Jolia; it aligns volumes and reports at the level of *concepts* rather than at the global level alone. A concept is a single semantic unit of localization; here, concepts are the $K = 102$ entries of a chest/abdomen anatomical taxonomy, typically a single organ but in some cases a compartment grouping several organs. We summarise the four building blocks below; see additional details in the Jolia publication [10].

Image encoder. The encoder is a 3D Atlas transformer [1] (22 M parameters, $6 \times 6 \times 6$ patches). It takes a CT volume resampled to 1.5 mm isotropic resolution and cropped to $192 \times 192 \times 192$ voxels, with an 11-channel multi-window CT representation (e.g., lung, mediastinum, soft-tissue, bone, and brain windows), and produces a multi-scale sequence of patch tokens at three feature resolutions ($d = 192$ per scale) together with a global `[CLS]` token obtained by max-pooling and concatenating patch tokens across the three scales, giving $d_{\texttt{[CLS]}} = 576$.

Concept-level cross-attention pooling. To extract a per-concept visual representation we introduce $K = 102$ learnable query tokens, one per concept. At each scale $s \in \{1, 2, 3\}$ we instantiate its own queries $Q^{(s)} = \{q_k^{(s)}\}$ and apply an independent cross-attention module that first LayerNorms the queries and the patch tokens and then performs a 1-head cross-attention of the queries onto the patches at that scale. The per-scale outputs are concatenated into the final per-concept representation

$$c_k = \left[\, \text{CrossAttn}\langle s\rangle\big(q_k^{(s)},\, \mathbf{z}\langle s\rangle\big) \,\right]_{s=1}^{3} \in \mathbb{R}^{576}, \tag{1}$$

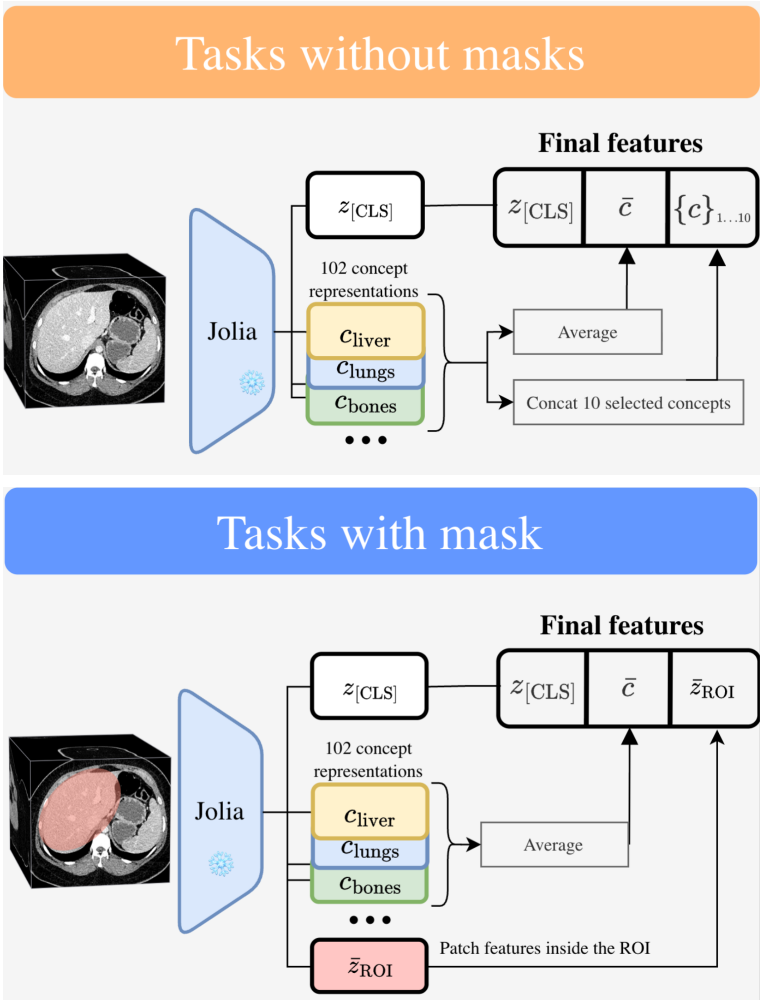

Fig. 2. Feature extraction for the two cases. **Top (no mask):** the feature concatenates the global $z_{\texttt{[CLS]}}$, the average of the 102 concept queries $\bar{c}$, and the 10 task-relevant organ queries $\{c\}_{1\ldots10}$. **Bottom (ROI target with a foreground mask):** the 10 organ queries are replaced by a mask-pooled patch feature $\bar{z}_{\text{ROI}}$ (`cls_roi`). The backbone is frozen; only the linear probe is trained.

where $\mathbf{z}\langle s\rangle$ is the patch-token set at scale s. Each query learns where to attend through the per-concept contrastive loss alone, with no spatial supervision; Fig. 3 shows that the queries nonetheless localize their organ. The pooling stack adds only $\sim 0.6\,\mathrm{M}$ parameters ($\sim 3\%$ of Atlas).

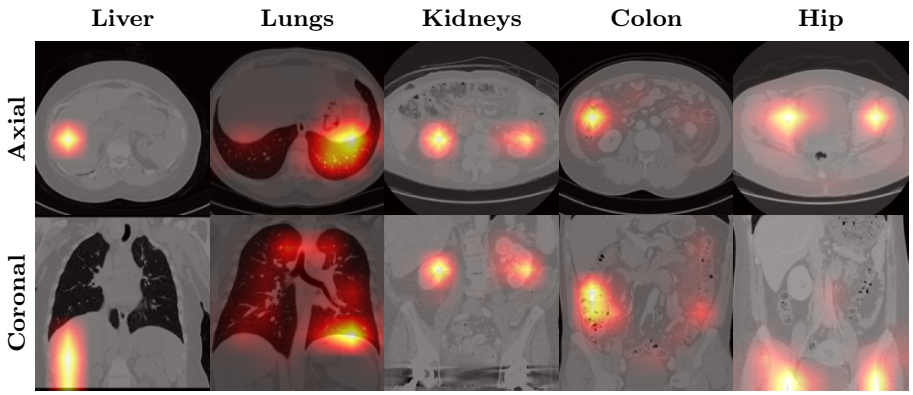

Fig. 3. Learned organ localization. Cross-attention maps of five organ queries, a by-product of ConQuer's concept-level alignment: each query attends to its organ without any spatial supervision at pretraining.

Text encoder. Following Pillar-0 [2], we use a frozen pretrained Qwen3-Embedding-8B [16] as text encoder. Each per-concept report section is encoded independently, and the full report is additionally encoded for the global loss; the 4096-d embeddings are projected to 576-d by a single learnable linear layer.

Loss. The per-concept contrastive loss is a symmetric InfoNCE applied independently per concept, combined with a global CLIP loss between $z_{\texttt{[CLS]}}$ and the full-report embedding t:

$$\mathcal{L}_{\text{ConQuer}} = \frac{1}{|\mathcal{K}_b|}\sum_{k\in\mathcal{K}_b}\mathcal{L}_{\text{CLIP}}^{(k)}, \qquad \mathcal{L} = \lambda_{\text{CLIP}}\,\mathcal{L}_{\text{CLIP}}^{\texttt{[CLS]}} + \lambda_{\text{ConQuer}}\,\mathcal{L}_{\text{ConQuer}}, \quad (2)$$

with $\lambda_{\text{CLIP}} = 1.0$ and $\lambda_{\text{ConQuer}} = 0.5$, a learnable per-concept temperature, and a concept-presence mask that drops (i, k) pairs whose report does not mention concept k.

2.2 Volume / patch sampling

Jolia processes *whole volumes*: the 192^3 input fits in memory, so there is no patch-based aggregation. During pretraining we apply `Rotation3D` (random 3D rotation), `RandomCrop3D` (192^3 out of 224^3), `ApplyWindowing` (deterministic, 11 windows), and `RandomIntensityShift` (mild intensity jitter). At linear-probing inference these are replaced by the deterministic eval transform of Section 3.2.

2.3 Anatomy-anchored feature recipe

The frozen backbone exposes several complementary cues (Table 1) that we concatenate into a single "+"-separated feature vector; the LP head's z-scoring and learned linear weights then pick the discriminative subset. Since the extractor is called without the target name, it cannot select a single organ-specific feature; we therefore emit a fixed feature and let the LP head specialise per target:

- targets without a mask (the 4 non-ROI abdominal findings, COVID, LUNA): `cls + queries_pooled` plus the 10 organ queries that cover every organ in the 17-target taxonomy (Table 2), a 6912-d vector;
- the 11 ROI targets, which ship a foreground organ mask: `cls + queries_pooled + cls_roi`, a 1728-d vector.

Because the head is trained per target (and the target organ is known at LP training time, from the labels), it learns to weight the relevant organ query, or the mask-pooled feature, without ever needing the target name at extraction. We use no per-task recipe selection at inference.

Table 1. Feature components exposed by the frozen Jolia backbone, concatenated to form the extracted feature.

Component	Description	Dim
`cls`	Multi-scale max-pool concat of patch tokens (vanilla forward)	576
`queries_pooled`	Mean over the 102 organ-query cross-attentions, $\times 3$ scales	576
`queries_slot_k`	A specific organ query (slot k), $\times 3$ scales	576
`cls_roi`	Mask-aware patch-token max-pool over foreground voxels, $\times 3$ scales	576
`cls_bbox10/20`	Volume re-cropped to mask 3D bbox +10/20% padding, re-forwarded	576

Table 2. The 10 organ queries used by the no-mask feature, and the validation targets each covers. Slot indices are positions in the 102-organ pretraining vocabulary.

Slot	Organ	Target(s) covered
2	adrenal glands	adrenal hyperplasia
4	aorta	atherosclerosis
20	colon	colorectal cancer
29	gallbladder	cholecystitis, gallstone
44	kidneys	hydronephrosis, kidney stone, renal cyst
46	liver	fatty liver, liver calcifications, liver lesion, liver cyst
47	lungs	covid, lung nodule malignancy
48	lymph nodes	lymphadenopathy
60	peritoneum	ascites
84	spleen	splenomegaly

2.4 Classification head and loss

For adaptation we use the **challenge's provided linear-probing pipeline without any changes**. It trains a **single linear layer** over a z-scored feature vector (train-set statistics), with a 2-class logit output, using **class-weighted cross-entropy** (`CrossEntropyLoss`, weights $w_c \propto 1/n_c$) – the only class-imbalance handling, with no oversampling. The optimizer is AdamW (weight decay 10^{-4}, no warm-up, constant learning rate); the pipeline sweeps 13 learning rates on a log grid from 10^{-5} to 10^{-1}, trains each for up to 1000 epochs with patience-50 early stopping on validation balanced accuracy, and keeps the best learning rate per target.

Jolia handles large 3D volumes without sliding-window inference: each volume is resampled and center-cropped to 192^3, which fits on a single GPU. For the 11 ROI targets the provided organ mask drives the `cls_roi` pooling; the other targets rely on the organ queries, which localize anatomy without any mask.

2.5 Post-processing

We apply no post-processing beyond selecting the LP head from the best learning rate on validation: no test-time augmentation, no checkpoint ensembling, and no per-task recipe selection. Feature extraction needs a single forward per volume, keeping per-case runtime under 5 s. Table 3 shows how the container picks the feature from the inputs it receives: the official runner provides masks but no target name, so every case goes through the mask / no-mask branches above.

Table 3. Feature selected by the container from the inputs it receives. `y_hat` is the extracted feature handed to the LP head.

Inputs	Feature extracted	`y_hat` dim
target name given	per-target oracle recipe (not used by the challenge)	576–1728
mask, no target name	`cls` + `queries_pooled` + `cls_roi`	1728
neither	`cls` + `queries_pooled` + 10 organ queries	6912

3 Experiments

3.1 Dataset and evaluation metrics

The challenge provides a large-scale CT pretraining dataset and a downstream training & validation set of **17 binary classification targets across 3 datasets**, covering abdomen and chest. We submit to **Task 1 (Linear Probing), All-Data Track** only; we do not submit the Coreset track or Task 2. Table 4 summarizes the validation data.

The **classification metrics** are Balanced Accuracy (BalAcc, the mean of sensitivity and specificity) and AUROC. The **efficiency metric** is running time per case (seconds), measured including Docker startup; for Task 1 it is the average feature-extraction time per target.

Table 4. Validation data for the 17 binary targets of Task 1.

Dataset	Region	# diseases	# val volumes	Notes
AMOS-clf	Abdomen	15	1463	11 ROI diseases ship foreground masks
COVID-CT	Chest	1	372	binary COVID vs. not
LUNA25	Chest	1	1200	lung nodule malignancy

3.2 Implementation details

Preprocessing. Volumes are loaded with `nibabel` and reoriented to canonical RAS+, matching Jolia's training pipeline. The deterministic eval transform applied to every volume is:

1. Reorient to RAS+ via `nibabel`.
2. `PrepareVolume(depth_last=True, flip_depth=True)` – move depth axis last, flip depth.
3. `Resample3D(1.5 mm)` isotropic.
4. `Crop3D(`224^3`)` center crop.
5. `Pad3D(`224^3`)` if needed.
6. `CenterCrop3D(`192^3`)` (replaces `RandomCrop3D` at inference).
7. 11-channel CT windowing.

For ROI diseases the mask receives the same spatial pipeline (steps 1–6) but is never windowed. For fast data loading on the large pretraining set we cache the text encoder outputs once per dataset and read volumes from a sharded WebDataset.

Environment settings. The development and evaluation environment is given in Table 5.

Table 5. Development environments and requirements. (mandatory table)

System	Ubuntu 22.04 LTS
CPU	AMD EPYC / Intel Xeon (lab cluster)
RAM	$\geq$ 32 GB (container capped at 32 GB)
GPU (number and type)	1× NVIDIA RTX 4090 24 GB (inference); pretraining: 8× NVIDIA H100 80 GB
CUDA version	12.4
Programming language	Python 3.11
Deep learning framework	PyTorch 2.5.1 + cuDNN 9

Training protocols. (1) Data augmentation. None at LP training time: features are extracted once with the deterministic eval transform and cached,

then the LP head is trained on z-scored vectors. Augmentation (Section 2.2) is used only during pretraining. **(2) Data sampling.** Class imbalance is handled by weighted CE, not by resampling; metrics are computed on a class-balanced subset of the validation split per target. **(3) Optimal model selection.** Highest validation balanced accuracy across the 13-LR sweep; one LP head per target. Table 6 lists the protocol.

Table 6. Training protocols for Task 1 – Linear Probing. The linear-head rows are the challenge's provided LP pipeline, used unchanged. (mandatory table)

Pretrained backbone (if any)	3D Atlas transformer (22 M), Jolia run, step 120,000
Backbone tuning strategy	**Frozen**
Batch size (pretraining)	48 (global, 8× H100)
Patch / volume size	$192 \times 192 \times 192$ at 1.5 mm, patch 6^3
Total pretraining epochs	120,000 steps
Linear-head epochs	up to 1000, patience-50 early stop on val BalAcc
Optimizer	AdamW (weight decay 10^{-4})
Initial learning rate (lr)	best of 13 LRs in $[10^{-5}, 10^{-1}]$ log grid
Lr decay schedule	constant (no decay)
Pretraining time	∼3 days on 8× H100
Linear-head training time	∼2–3 min / disease / LR on 1× RTX 4090
Loss function	class-weighted cross-entropy ($w_c \propto 1/n_c$)
Number of model parameters	21.8 M (Atlas) + 0.6 M (queries) = 22.4 M frozen; LP head = 2×feature-dim (≈3.5 k for the 1728-d ROI feature, ≈13.8 k for the 6912-d feature)[4]
Number of FLOPs	≈497 GMACs/volume (`fvcore` total; backbone 492 + organ queries 5)[5]

3.3 Quantitative results on the validation set

Table 7 compares Jolia with the eight official challenge baselines (Codabench [15]). On chest, Jolia is the strongest entry after SPECTRE and Pillar-0 (BalAcc 0.8447). On abdomen and overall it sits mid-pack: its overall BalAcc (0.7575) trails SPECTRE (0.7715), Curia (0.7711) and Pillar-0 (0.7683) but is ahead of Merlin, 3DINO, VoCo, CT-FM and CT-NEXUS. We report this honestly: Jolia's edge is on chest, while the small organ-localized abdominal findings are where it loses ground (Section 3.6). Efficiency is a strength: Jolia's frozen encoder has only 22.4 M parameters and extracts features in ∼5 s/case, well under the 40 s budget.

Ablation: feature choice. Table 8 reports the mean BalAcc per region group for each candidate feature. The deployed feature uses `cls` + `queries_pooled` + `cls_roi` on the 11 ROI targets and `cls` + `queries_pooled` + 10 organ queries on the rest

Table 7. Task 1 – Linear Probing, **All-Data Track**: mean BalAcc / AUROC per region over the 17 validation targets. Baselines are the challenge's official numbers (Codabench [15]); Jolia uses the deployed feature. Best per column in **bold**. (mandatory table)

	Chest (2)		Abdomen (15)		Overall (17)			
Method	BalAcc	AUROC	BalAcc	AUROC	BalAcc	AUROC	Time (s)	# Params (M)
CT-NEXUS	0.6040	0.5700	0.6241	0.6177	0.6217	0.6121	–	–
VoCo	0.6495	0.6345	0.6955	0.6906	0.6901	0.6840	–	–
CT-FM	0.6380	0.6100	0.6991	0.6995	0.6919	0.6889	–	–
3DINO	0.7009	0.7046	0.7216	0.7403	0.7192	0.7361	–	–
Merlin	0.6905	0.7090	0.7287	0.7303	0.7242	0.7278	–	–
Pillar-0	0.8695	**0.8965**	0.7548	0.7838	0.7683	**0.7971**	–	–
SPECTRE	**0.8700**	0.8838	0.7584	0.7713	**0.7715**	0.7845	–	–
Curia	0.7420	0.7655	**0.7749**	**0.7903**	0.7711	0.7874	∼4	86
Jolia (ours)	0.8447	0.8696	0.7459	0.7504	0.7575	0.7645	∼5	**22.4**

– the two configurations actually reachable without the target name. The organ-query and mask-pooling cues both help over the global `[CLS]` alone, and the abdominal headroom (ROI-11 / nonROI-4) is small across all variants.

Table 8. Ablation – mean BalAcc per region group (ROI-11 / nonROI-4 / COVID / LUNA), each feature applied uniformly within the group. "10 slots" is the no-mask deployed feature; rows with `cls_roi` use the foreground mask. "–" = not applicable.

Feature	ROI-11	nonROI-4	COVID	LUNA
`cls`	0.7068	0.7564	0.9844	0.6850
`queries_pooled`	0.7248	0.7444	0.9844	0.6600
`cls + queries_pooled`	0.7193	0.7622	0.9844	0.6700
`queries_slot`	0.7454	0.7262	0.9844	0.6950
`cls + queries_slot`	0.7323	0.7542	0.9844	0.6850
`queries_slot + queries_pooled`	0.7357	0.7683	0.9844	0.6800
`cls + queries_pooled +` 10 slots	–	0.7589	0.9844	0.7050
`cls_roi`	0.7289	0.7564	0.9844	0.6850
`cls_roi + queries_pooled`	0.7361	0.7622	0.9844	0.6700
`cls + cls_roi + queries_pooled`	0.7412	0.7622	0.9844	0.6700
`cls_bbox20 + cls_roi + queries_pooled`	0.7312	0.7622	0.9844	0.6700

Per-disease results. Table 9 breaks the comparison with Curia (the closest baseline overall) down per disease. Jolia wins both chest targets and four abdominal ones (ascites, cholecystitis, colorectal cancer, kidney stone); Curia wins the others (liver lesion ties). Jolia's deficit concentrates on small organ-localized findings (liver cyst, calcifications, fatty liver, lymphadenopathy).

3.4 Qualitative analysis on the validation set

Fig. 4 shows, for each region, one correctly classified and one misclassified validation case, with the predicted positive probability. *Chest.* A clear COVID study

Table 9. Per-disease results on the **validation** split. Curia is the official challenge baseline (Codabench [15]); Jolia uses the deployed feature. BalAcc winner per disease in bold.

Disease	Curia BAcc	Curia AUC	Jolia BAcc	Jolia AUC
adrenal hyperplasia	**0.7860**	0.8200	0.6857	0.6776
ascites	0.7500	0.7780	**0.7941**	0.7850
atherosclerosis	**0.8090**	0.8090	0.7500	0.7059
cholecystitis	0.7550	0.7670	**0.7843**	0.7493
colorectal cancer	0.7500	0.7320	**0.7692**	0.6879
fatty liver	**0.9490**	0.9780	0.8590	0.9034
gallstone	**0.8190**	0.8560	0.7586	0.8133
hydronephrosis	**0.9060**	0.9520	0.8750	0.9141
kidney stone	0.6770	0.6780	**0.6899**	0.7177
liver calcifications	**0.6670**	0.6220	0.6389	0.6065
liver cyst	**0.6680**	0.6820	0.6230	0.6545
liver lesion	0.6940	0.7510	0.6940	0.7166
lymphadenopathy	**0.8060**	0.8040	0.7222	0.7238
renal cyst	**0.6880**	0.6990	0.6729	0.7137
splenomegaly	**0.9000**	0.9260	0.8714	0.8873
AMOS-15 mean	**0.7749**	0.7903	0.7459	0.7504
covid	0.8440	0.8560	**0.9844**	0.9922
lung nodule malignancy	0.6400	0.6750	**0.7050**	0.7471
Overall mean (17)	**0.7711**	0.7874	0.7575	0.7645

is detected with $p = 1.00$; the single COVID error is a subtle case ($p = 0.24$) with only faint ground-glass, near the saturated decision boundary of this split. *Abdomen.* An enlarged spleen (outlined) is correctly flagged with $p = 1.00$; the hard case is a small liver cyst (outlined) missed with $p = 0.06$ – the lesion is only a few voxels wide, below the resolution at which the mask-pooled feature carries a useful signal, which is the typical failure mode behind Jolia's lower abdominal scores.

3.5 Results on the final testing set

This is a placeholder. We will add the testing results during the revision phase, after they are announced at CVPR 2026.

3.6 Limitation and future work

Abdominal organ-localized findings are where Jolia trails Curia. **Resolution is the main limit**: Jolia always encodes the whole volume at a single 1.5 mm, 192^3 scale and never sees a zoomed-in crop, so small targets (liver/renal cysts, calcifications, fatty liver, lymphadenopathy) are only a few voxels wide and are washed out before they reach a token. Re-running Jolia on a cropped, higher-resolution view of a candidate region – so it can "zoom in" rather than relying on a single global pass – is the natural way to recover this signal, and

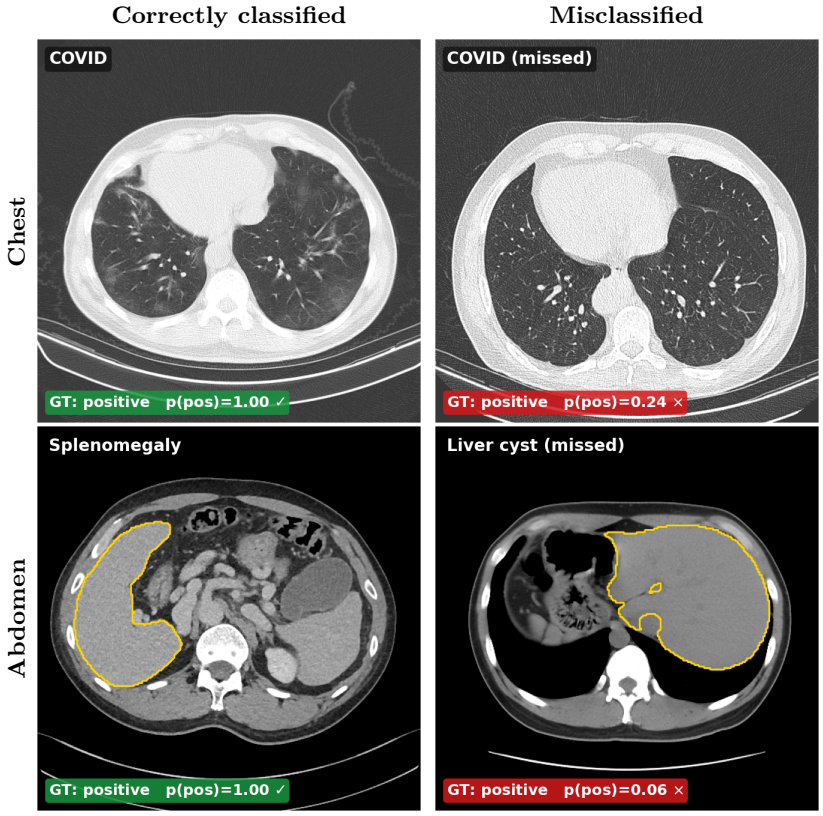

Fig. 4. Qualitative validation cases. One correctly classified (left) and one misclassified (right) case per region, annotated with the ground truth and the predicted positive probability. Soft-tissue window for abdomen, lung window for chest; the organ ROI is outlined in yellow when a mask is available.

our main direction for future work. **The AMOS-15 validation split is small per disease** ($\sim$ 70–100 cases), so per-disease BalAcc has wide variance; the final test set will give a more reliable ranking. Finally, **we did not submit Task 2 or the Coreset track** by choice, focusing on the All-Data Track.

4 Conclusion

ConQuer's concept-level contrastive pretraining yields, in Jolia, 102 anatomy-anchored concept queries alongside the global `[CLS]` token. With a fixed feature and a linear probe that selects the relevant signal per target, Jolia reaches BalAcc 0.7575 / AUROC 0.7645: it beats the Curia 2D baseline on chest (0.8447 vs. 0.7420), trails it on abdomen and overall (Curia 0.7711), and ranks mid-pack among the eight official baselines, the gap concentrating on small organ-localized abdominal findings that fall below our single 192^3 input resolution. Re-running Jolia on zoomed-in, higher-resolution crops to recover these small findings is our main direction for future work. The system ships as a < 4 GB Docker container running in < 5 s/case, well within the challenge's 40 s/case budget.

Acknowledgements We thank all the data owners for making the medical images publicly available and CodaLab [15] for hosting the challenge platform.

Disclosure of Interests. The authors are affiliated with Raidium.

References

1. Agrawal, K.K., Lian, L., Liu, L., Harguindeguy, N., Li, B., Bick, A.G., Chung, M., Darrell, T., Yala, A.: Atlas: Multi-scale attention improves long context image modeling. ArXiv **abs/2503.12355** (2025), https://api.semanticscholar.org/CorpusID:277066076 3
2. Agrawal, K.K., et al.: Pillar-0: A new frontier for radiology foundation models. arXiv preprint arXiv:2511.17803 (2025) 2, 5
3. Blankemeier, L., et al.: Merlin: A vision language foundation model for 3D computed tomography. Research Square (2024), rs-3 2
4. Claessens, C., et al.: Scaling self-supervised and cross-modal pretraining for volumetric CT transformers. arXiv preprint arXiv:2511.17209 (2025) 2
5. Dancette, C., et al.: Curia: A multi-modal foundation model for radiology. arXiv preprint arXiv:2509.06830 (2025) 2
6. Ericsson, L., Gouk, H., Hospedales, T.M.: How well do self-supervised models transfer? CVPR (2021) 2
7. Hamamci, I.E., et al.: A foundation model utilizing chest CT volumes and radiology reports for supervised-level zero-shot detection of abnormalities. CoRR (2024) 2
8. Houlsby, N., Giurgiu, A., Jastrzebski, S., et al.: Parameter-efficient transfer learning for NLP. In: ICML (2019) 2
9. Hu, E.J., Shen, Y., Wallis, P., et al.: LoRA: Low-rank adaptation of large language models. In: ICLR (2022) 2

10. Khlaut, J., Corbière, C., Callard, B., Prat, A., Butsanets, L., Saporta, A., Danielou, T., Machado, L., Le Floch, K., Boeken, T., Manceron, P., Dancette, C.: Concept-level vision-language alignment for 3D CT contrastive learning. arXiv preprint (2026), preprint; **TODO** add arXiv id / venue 2, 3
11. Pai, S., et al.: Vision foundation models for computed tomography. arXiv preprint arXiv:2501.09001 (2025) 2
12. Shui, Z., Zhang, J., Cao, W., et al.: Large-scale and fine-grained vision-language pre-training for enhanced ct image understanding. International Conference on Learning Representations (ICLR) (2025) 2
13. Wu, L., Zhuang, J., Chen, H.: Large-scale 3D medical image pre-training with geometric context priors. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025) 2
14. Xu, T., et al.: A generalizable 3D framework and model for self-supervised learning in medical imaging. npj Digital Medicine **8**(1), 639 (2025) 2
15. Xu, Z., Escalera, S., Pavao, A., et al.: Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform. Patterns **3**(7), 100543 (2022) 9, 10, 11, 13
16. Zhang, Y., Li, M., Long, D., et al.: Qwen3 embedding: Advancing text embedding and reranking through foundation models. arXiv preprint arXiv:2506.05176 (2025) 5

Table 10. Checklist Table. Please fill out this checklist table in the answer column. (**Delete this Table in the camera-ready submission**)

Requirements	Answer
A meaningful title	Yes
The number of authors (≤6)	6
Author affiliations and ORCID	Yes (all 6 authors)
Corresponding author email is presented	Yes (julien.khlaut@raidium.eu)
Task and track (LP All / LP Coreset / EAO VO / EAO VL) are clearly stated in the abstract	Yes (LP All-Data)
Validation BalAcc and AUROC are presented in the abstract	Yes (0.7575 / 0.7645)
Introduction includes at least three parts: background, related work, and motivation	Yes
A pipeline/network figure is provided	Fig. 1
Pre-processing	Sec. 3.2
Strategies for data augmentation	Sec. 2.2
Strategy for class imbalance	Sec. 2.4
Strategies to improve model inference (under 40s/case)	Sec. 2.5
Post-processing	Sec. 2.5
Environment setting table is provided	Table 5
Training protocol table is provided (Task 1 and/or Task 2)	Table 6
Ablation study	Table 8
Efficiency evaluation results (running time) are provided	Sec. 3.3
Qualitative example (success and failure) is provided	Fig. 4
Limitation and future work are presented	Yes
Reference format is consistent	Yes (splncs04)
Main text ≥ 8 pages (not including references and appendix)	Yes