REFORM-3D

A Representation-Centric Evaluation Framework for 3D Medical Vision Foundation Models

Maregu Assefa[1] · Muzammal Naseer[1,2] · Divya Velayudhan[1]
Gedamu Alemu Kumie[1] · Naoufel Werghi[1]

[1] Department of Computer Science, Khalifa University, Abu Dhabi, UAE
[2] University of Western Australia

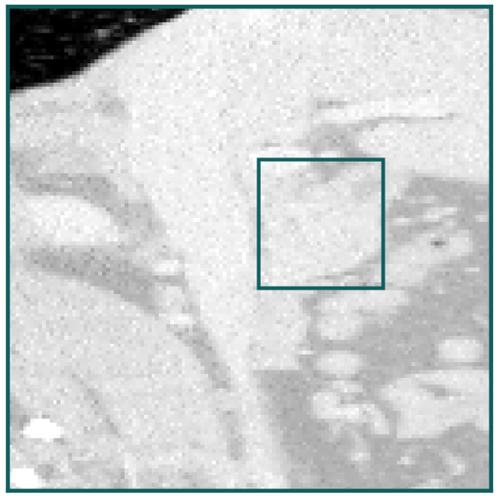
Spacing sensitivity

CT–MR retrieval

Anatomical transfer

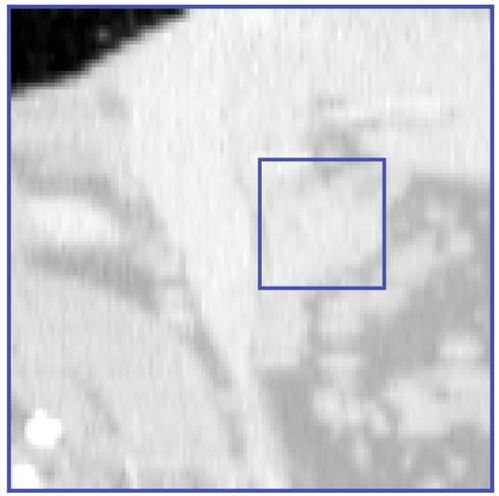
A downstream score mixes representation and adaptation

Spacing can change image statistics while the underlying scan stays the same.



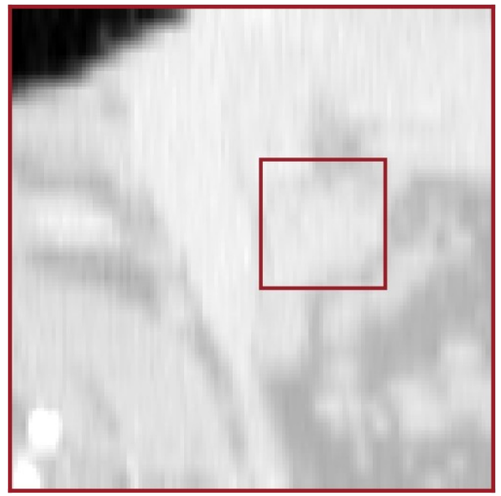
Reference

1 × 1 × 1 mm



Moderate

1 × 1 × 3 mm



Severe

1 × 1 × 5 mm

What remains stable—and what anatomical information remains usable?

Three questions about the same frozen encoder

A profile of complementary properties, without a pooled overall score.

Acquisition Robustness

Same scan, different spacing

Drift ↓ · CKA ↑
Semantic transfer ↑

Cross-Modal Anatomical Alignment

Same organ, CT ↔ MR

mAP ↑
Top-1 ↑ · Top-5 ↑

Anatomical Transfer

Limited support, new cases

Centroid BA · Probe BA
Adaptation gain

Stable embeddings and recoverable anatomy are different properties.

Med3DINO controls the recipe and varies geometric policy

62,638 unlabeled volumes · CT / MR / PET / CTA / CBCT · ViT-L/16

ISO

Isotropic normalization

1.5 × 1.5 × 1.5 mm
then voxel-relative crops

REL

Voxel-relative sampling

Native spacing
voxel-relative crop scales

SA

Spacing-aware resampling

Native spacing
physical through-plane extent
anisotropy-aware rotations

Held fixed: corpus · initialization policy · architecture · DINO/iBOT/KoLeo
non-geometric augmentations · crop counts/output sizes · 125K updates

Interpret the intervention as a complete geometric pretraining policy.

Each module uses its own evaluation surface

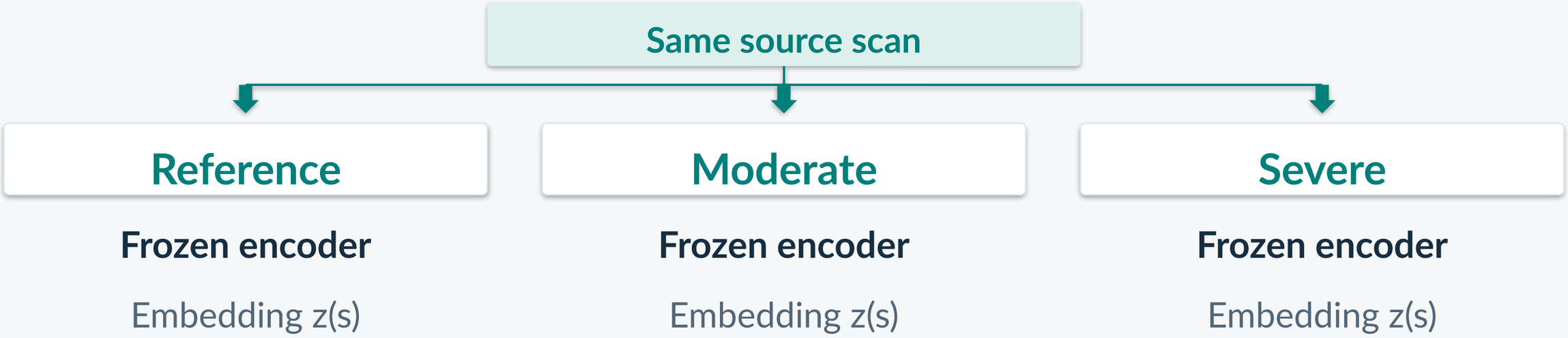
Nine external encoders contextualize the controlled ISO / REL / SA triplet.

Surface	Primary displayed datasets	Frozen readout
Controlled spacing	AbdomenAtlas · KiTS23 ImageCAS · TS-MR	CLS / GAP for 3DINO and Med3DINO
CT–MR retrieval	TotalSeg CT–MR · MMWHS AMOS22	Native interface for other encoders
Anatomical transfer	TotalSeg CT–MR · MMWHS AMOS22	Same support budget; case-disjoint queries

External encoders: BrainIAC · OpenMind · VoCo · 3DINO · SAM-Med3D
Hi-End-MAE · MambaMIM · MiM · GL-MAE

Change spacing; hold scan identity and labels fixed

Compare all three geometry variants of the same scan using frozen features.



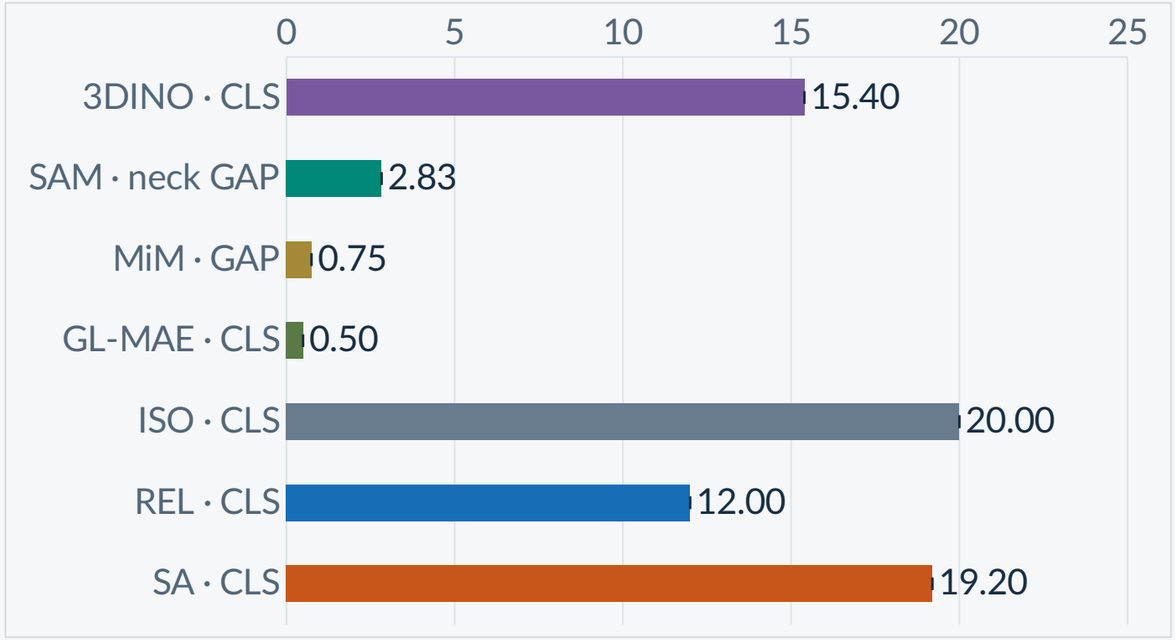
Drift ↓	CKA ↑	Semantic transfer ↑
Individual embedding movement	Centered kernel alignment: cohort feature geometry	Source-trained readout on target-spacing cases

Geometric stress testing includes resampling effects, not scanner physics.

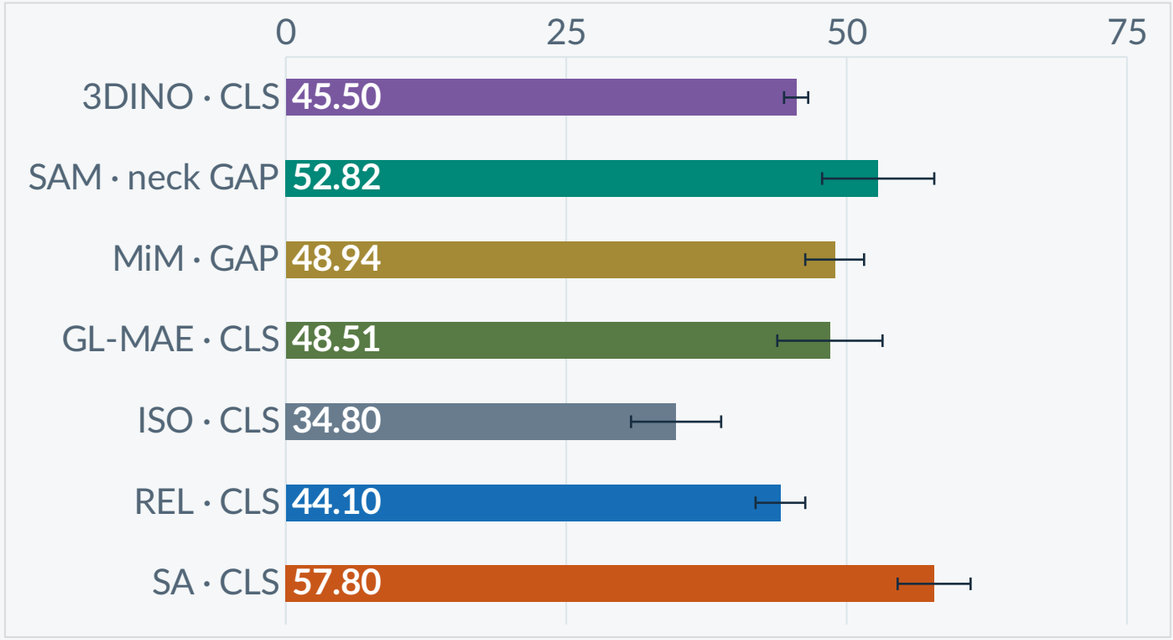
Low drift does not ensure stronger semantic transfer

KiTS23 · matched source/target cases · values on the manuscript ×100 scale

Drift ↓



Semantic transfer ↑



Whiskers: ± sample SD over 3 evaluation seeds; checkpoints fixed.

Within Med3DINO CLS: REL is more stable; SA transfers better on KiTS23.

Retrieve the same organ across CT and MR

Balanced organ pools · cosine ranking · equal weight for both retrieval directions

CT query

Organ crop
with retained context

Frozen representation

Balanced MR candidate pool

Same-organ positives and competing organs

Rank by cosine similarity

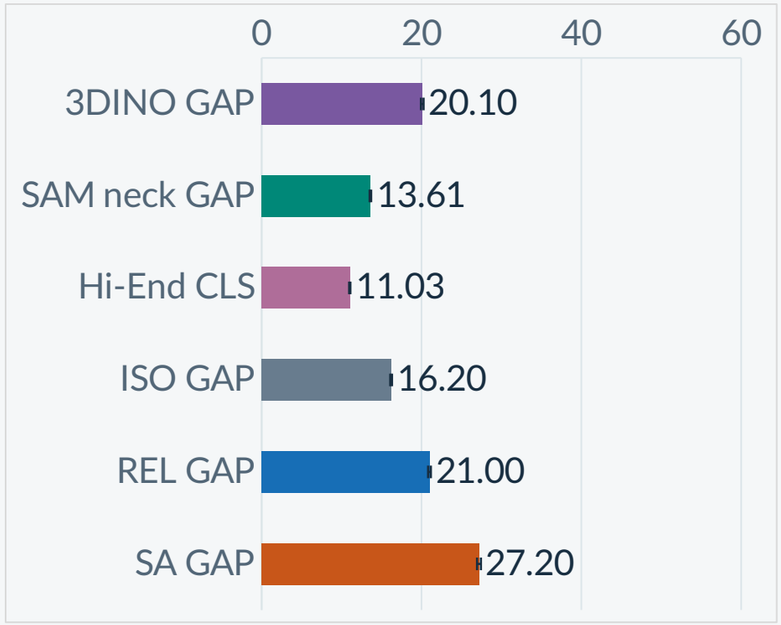
Repeat MR → CT and average directions equally.

Recoverable organ identity across modalities is the measured property.

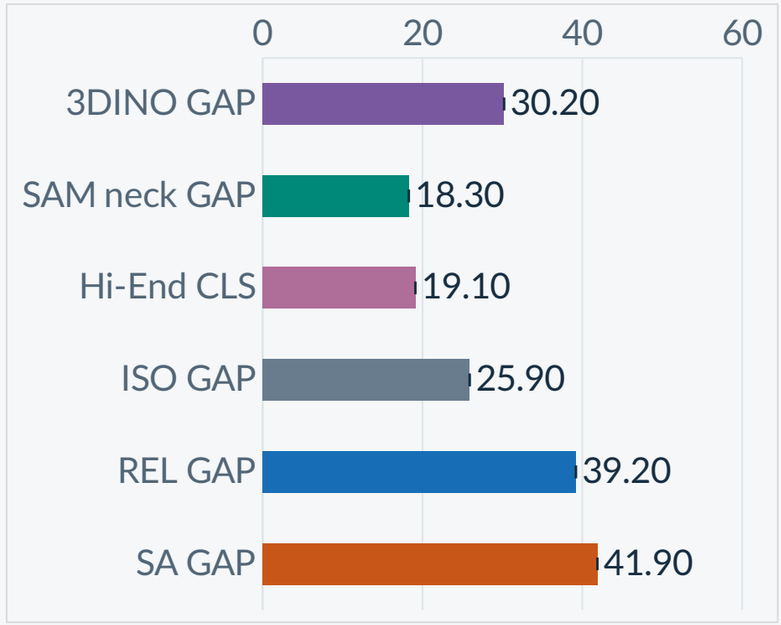
SA leads the controlled multi-organ retrieval comparison

Bidirectional mAP ↑ · primary context-inclusive crops · GAP for the Med3DINO triplet

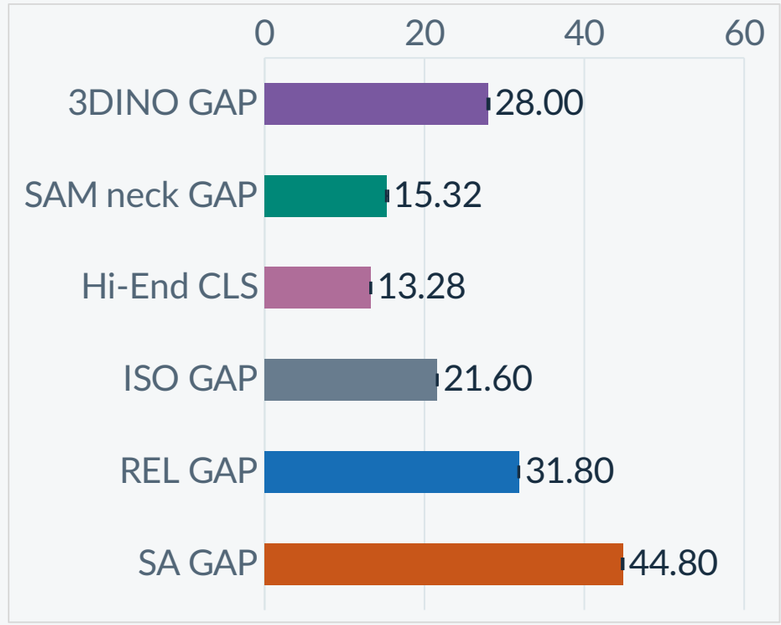
TotalSeg CT–MR



MMWHS



AMOS22



Values ×100. Whiskers: ± sample SD of evaluation seeds, not pretraining runs.

The controlled result supports organ recoverability in these crop constructions.

Use a little labeled support; test on different cases

Two support samples per organ and modality · frozen encoder throughout

Labeled support

Fit standardizer + class centroids / probe



Case-disjoint queries

Apply fitted readouts unchanged

Centroid BA	Probe BA	Adaptation gain
Nearest-class-mean separability	Limited-label recoverability	Probe BA − centroid BA computed within each seed

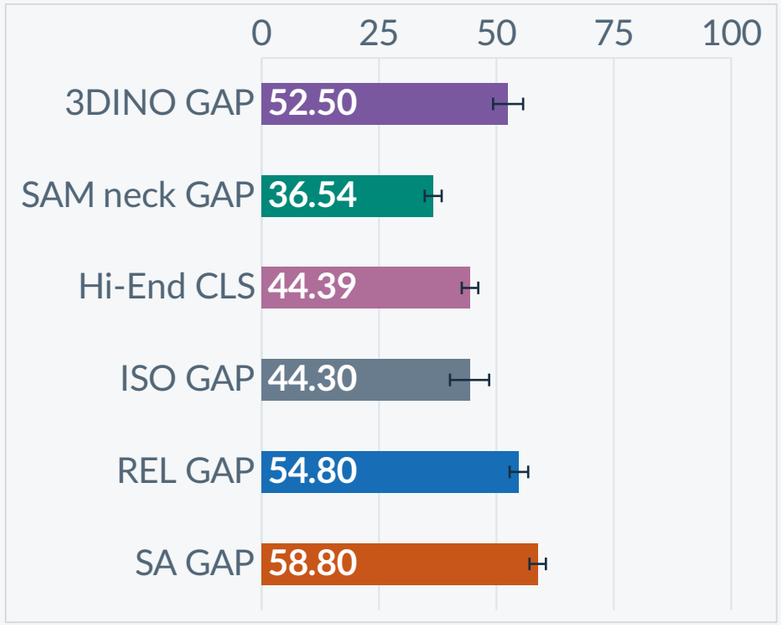
BA = mean recall across organ classes. Encoder weights stay frozen.

Separate final accuracy from the size of the adaptation gain.

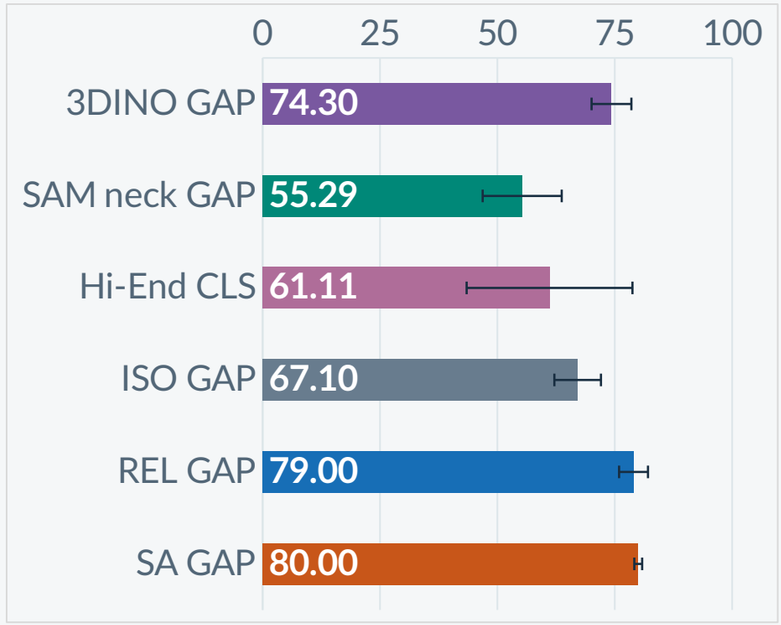
SA has the highest probe BA within Med3DINO

Case-disjoint organ classification · two supports / organ / modality · probe BA ↑

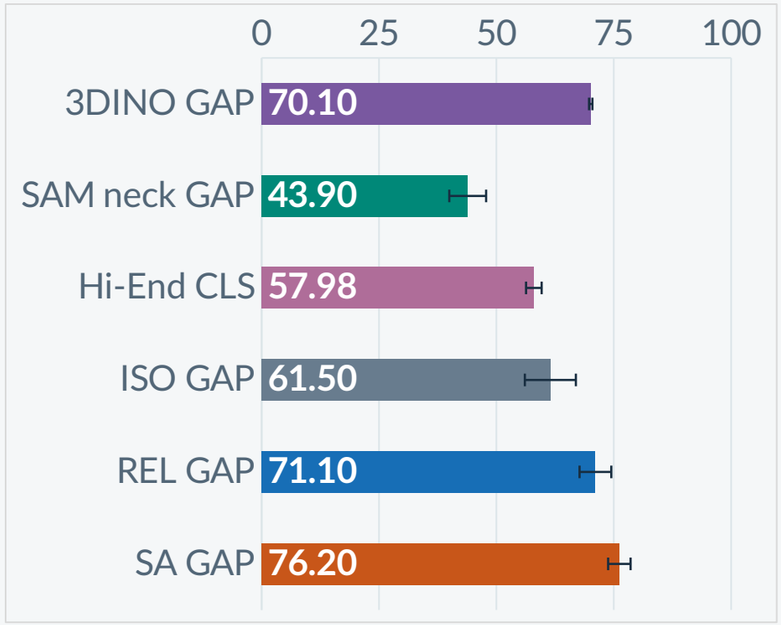
TotalSeg CT–MR



MMWHS



AMOS22



Values ×100. Whiskers: ± sample SD over 3 evaluation seeds. External inputs/interfaces vary.

Higher reported organ recoverability can coexist with greater spacing drift.

Downstream transfer adds practical, task-specific evidence

Frozen encoder + trained ViTAdapterUNETR decoder · manuscript segmentation summary

Dataset	3DINO	ISO	REL	SA
BTCV · seen	82.10 / 9.4	64.71 / 15.56	65.50 / 7.94	75.69 / 5.76
KiTS23 · unseen	80.75 / 13.87	88.00 / 19.61	88.11 / 14.57	88.50 / 9.15
AMOS-MR · unseen	80.69 / 5.54	82.12 / 4.69	82.07 / 5.58	82.17 / 4.83
TDSC-ABUS · US	47.70 / 89.10	72.78 / 75.61	74.94 / 68.72	75.91 / 65.29

Cell: Dice (%) ↑ / HD95 (mm) ↓. No seed SD is reported in this table.

Exposure labels refer to Med3DINO. US is the unseen-modality setting.

Task scores complement the frozen-feature profile; they do not replace it.

Use the profile to guide subsequent validation

The standalone repository provides a YAML check → run workflow.

Configure → check → run

Configure configs/my-kits23.yaml

reform3d check --config configs/my-kits23.yaml
reform3d run --config configs/my-kits23.yaml

Inspect reports and saved provenance

Interpret within the tested scope

Spacing is a geometric stress test

Retrieval retains anatomical context

Geometry results use one matched recipe

Validate the intended downstream task

Assess frozen features first; validate the intended application next.

Three properties, three questions, one usable framework

1 **Stability is one property**

Read drift alongside semantic and anatomical recoverability.

2 **Geometry and readout both matter**

Keep controlled policies, datasets and feature interfaces explicit.

3 **A profile supports better questions**

Use REFORM-3D before task-specific validation.

Project page

Code

Med3DINO weights

Project, code and checkpoint links

External encoders retain their designated feature interfaces

The controlled triplet and heterogeneous external methods answer different comparison questions.

Encoder	Architecture	Evaluated features	Pretraining objective
BrainIAC	ViT-B	First patch token	SimCLR
OpenMind	Primus-M	GAP	Volume contrastive
VoCo	SwinUNETR	Multiscale GAP	Volume contrastive
3DINO	ViT-L	CLS / GAP	DINOv2
SAM-Med3D	ViT-B	Post-neck GAP	Dice + CE
Hi-End-MAE	ViT	CLS	Hierarchical MAE
MambaMIM	CNN–Mamba	Final-stage GAP	Masked reconstruction
MiM	ConvViT-B	GAP	MAE + InfoNCE
GL-MAE	ViT-B	CLS	MAE + consistency

Med3DINO: ViT-L/16 · 96^3 evaluation inputs · CLS / GAP · shared DINOv2-based recipe

External strengths depend on the measured property

One dataset; native features; separate metrics. These columns are not a pooled score.

Encoder	Readout	mAP ↑	Centroid BA ↑	Probe BA ↑
BrainIAC	First patch	21.60 ± 0.00	35.30 ± 2.70	42.50 ± 2.82
OpenMind	GAP	17.60 ± 2.45	23.40 ± 3.67	24.60 ± 2.93
VoCo	Multiscale GAP	19.30 ± 0.00	29.20 ± 3.03	47.60 ± 7.34
3DINO	GAP	30.20 ± 0.00	65.90 ± 5.51	74.30 ± 4.24
SAM-Med3D	Post-neck GAP	18.30 ± 0.00	34.79 ± 1.88	55.29 ± 8.41
Hi-End-MAE	CLS	19.10 ± 0.00	44.05 ± 10.80	61.11 ± 17.66
MambaMIM	Final GAP	18.27 ± 0.00	37.70 ± 12.40	47.09 ± 8.66
MiM	GAP	18.26 ± 0.00	23.81 ± 3.83	46.16 ± 12.41
GL-MAE	CLS	17.40 ± 0.00	33.33 ± 14.59	42.33 ± 12.02

Values ×100; mean ± sample SD over seeds 42, 123, 456. No paired external significance claim.

CLS and GAP can change the interpretation

3DINO and Med3DINO · KiTS23 acquisition and MMWHS retrieval / transfer

Encoder / readout	KiTS drift ↓	KiTS CKA ↑	KiTS transfer ↑	MMWHS mAP ↑	MMWHS probe ↑
3DINO · CLS	15.40 ± 0.00	77.30 ± 0.00	45.50 ± 1.08	29.40 ± 0.00	61.60 ± 0.40
3DINO · GAP	3.70 ± 0.00	91.90 ± 0.00	39.90 ± 2.23	30.20 ± 0.00	74.30 ± 4.24
ISO · CLS	20.00 ± 0.00	73.60 ± 0.00	34.80 ± 4.02	28.10 ± 0.00	53.30 ± 7.20
ISO · GAP	3.90 ± 0.00	91.50 ± 0.00	43.80 ± 2.41	25.90 ± 0.00	67.10 ± 4.96
REL · CLS	12.00 ± 0.00	86.10 ± 0.00	44.10 ± 2.22	25.40 ± 0.00	58.10 ± 5.57
REL · GAP	4.40 ± 0.00	92.60 ± 0.00	41.60 ± 3.55	39.20 ± 0.00	79.00 ± 3.08
SA · CLS	19.20 ± 0.00	84.30 ± 0.00	57.80 ± 3.25	31.10 ± 0.00	62.80 ± 4.80
SA · GAP	5.80 ± 0.00	94.30 ± 0.00	49.20 ± 3.32	41.90 ± 0.00	80.00 ± 0.83

Mean ± evaluation-seed SD; all scores ×100. GAP averages the final normalized patch tokens.

Support-budget sensitivity is a separate evaluation

CLS probe BA · equal-weight means over TotalSeg CT–MR, MMWHS and AMOS22

Support / organ / modality	3DINO	REL	SA	SA highest mean	Queries: TS / MMWHS / AMOS
1	54.8	55.3	56.4	1/3	85/19/54
2	64.7	66.1	68.7	3/3	84/18/53
4	71.2	75.1	77.5	3/3	82/16/51
8	76.9	80.0	81.5	3/3	77–78/12/47

Query counts are per organ and modality. Reported means ×100; no SD supplied in this table.

Changing support also changes the query pool; interpret each budget separately.

Context changes scores and the size of the SA–REL gap

CLS retrieval · equal-weight AMOS22 and MMWHS means · five- or zero-voxel margins

Crop surface	ISO mAP	REL mAP	SA mAP	SA − REL [adjusted 95% CI]
M5 + context	15.3	33.2	42.7	+9.5 [8.6,10.4]
M0 + context	15.5	31.5	40.4	+8.8 [7.8,9.8]
M5 + organ-only	16.7	24.0	26.4	+2.4 [1.6,5.4]
M0 + organ-only	17.8	22.2	23.7	+1.5 [0.6,2.6]

Values ×100; differences in percentage points. Context retained / zeroed within original-mask crop boxes.

Stable ordering can coexist with strong context dependence.

Interpret the metric and its uncertainty together

Fixed checkpoints · evaluation seeds 42, 123 and 456 · scores and SDs ×100

Drift and CKA

Individual direction change vs centered cohort geometry

mAP and balanced accuracy

Equal retrieval-direction weights vs mean organ recall

Mean ± sample SD

Evaluation variability of fixed checkpoints
A reported 0.00 can reflect rounding.

Paired confidence intervals

Use the corresponding cohort, matched units and adjustment

Descriptive SD is not a significance test or a pretraining replication study.