FinAI: Fine-grained Alignment for Detail-Preserving Medical Vision-Language Pretraining

Jongsu Youn, Dongyoung Lee, JaeHyeong Bae, Sang-Il Choi, Sangtae Choi, Dae Ung Jo, Jongwon Choi

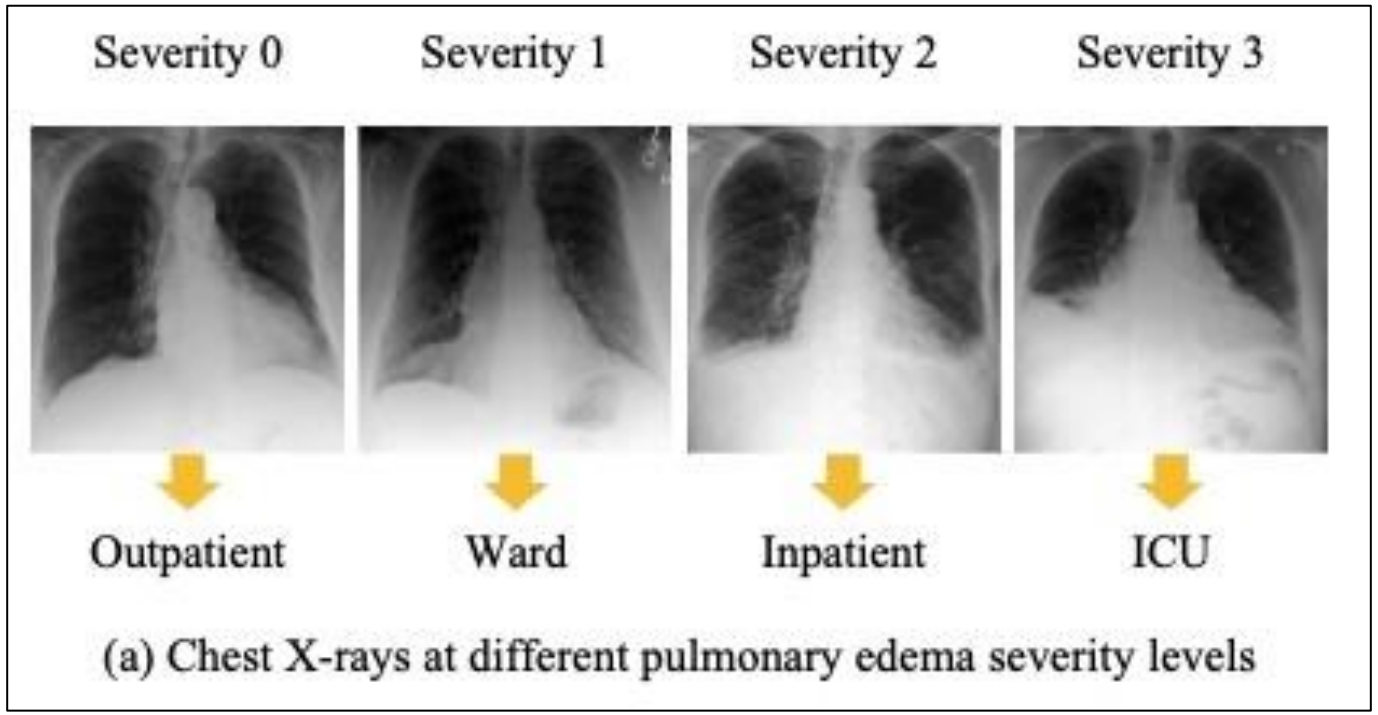








Motivation

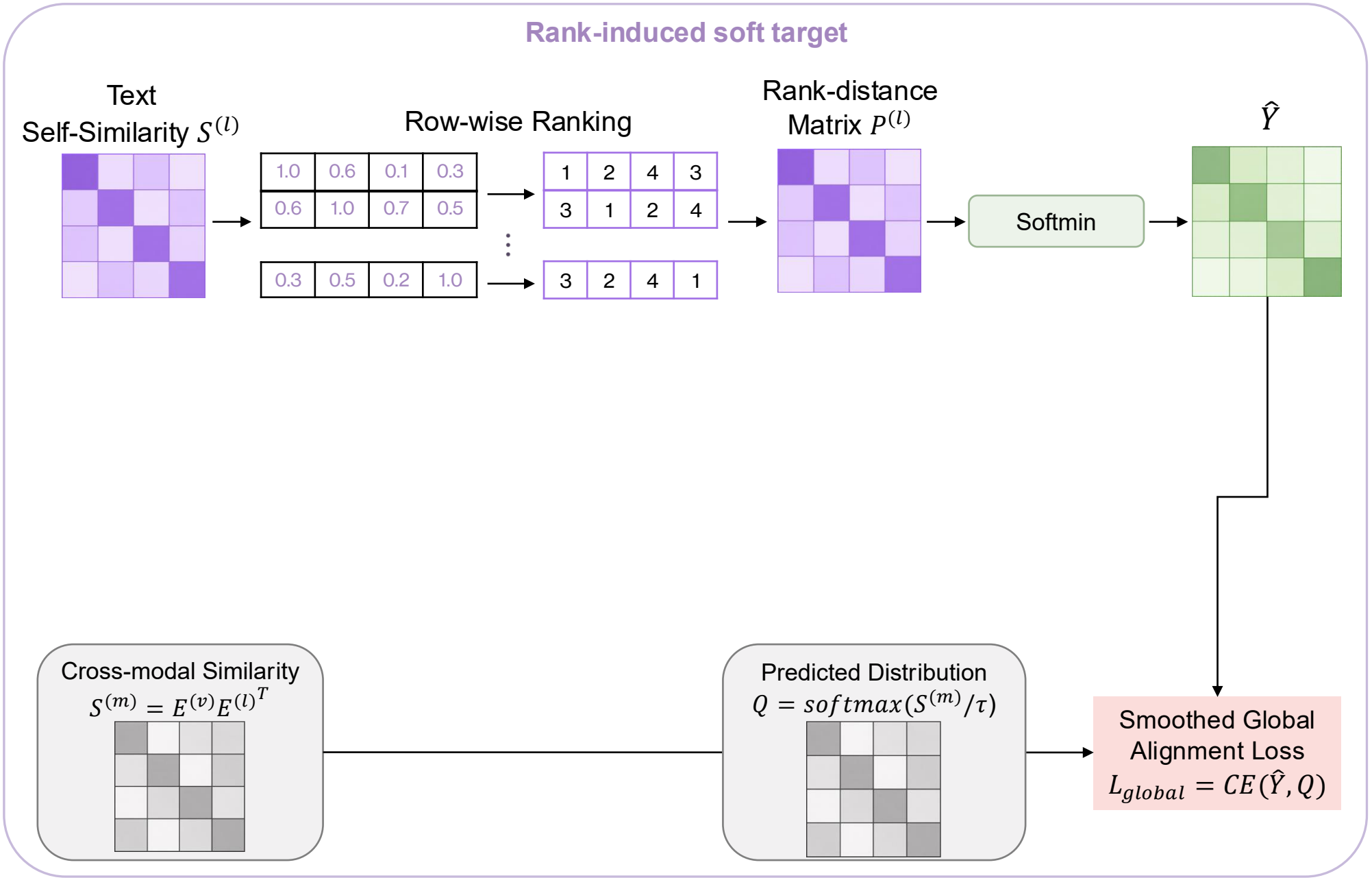


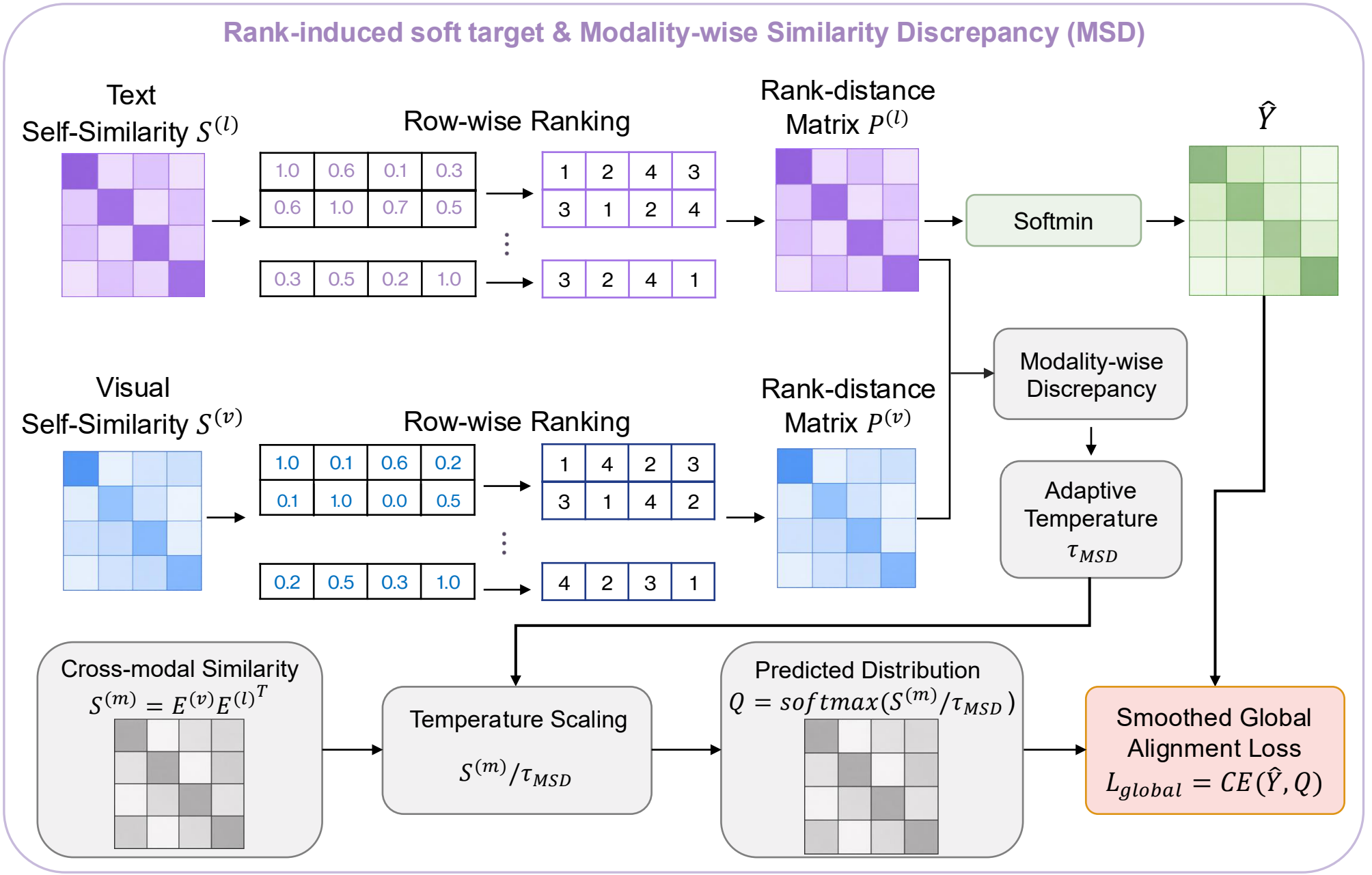
(a) Chest X-rays at different pulmonary edema severity levels

- Patients with the same disease may differ in severity, so representations should preserve **graded clinical differences**.

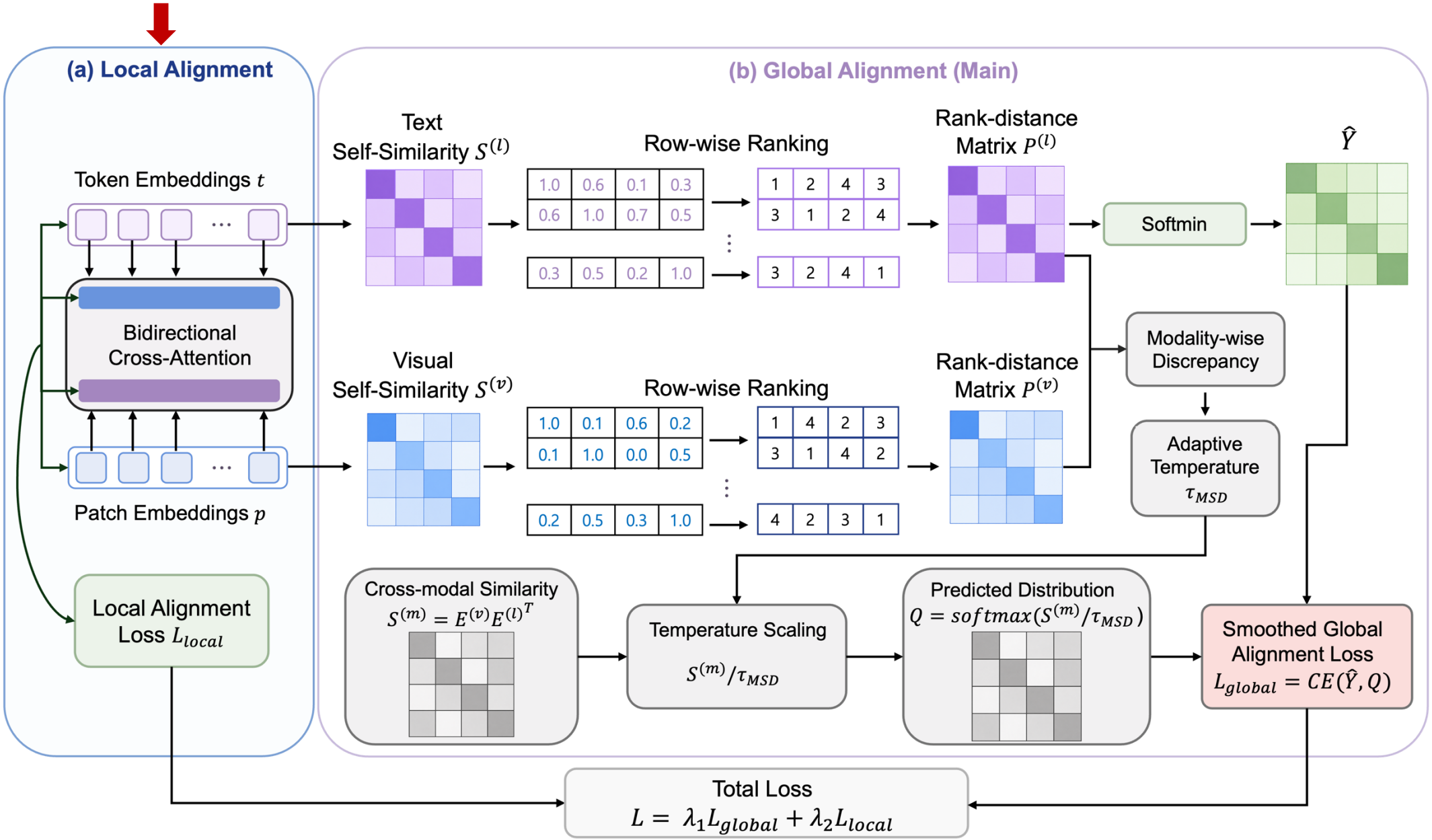
Motivation

- Hard one-hot targets do not distinguish degrees of clinical relatedness among non-paired samples.
- Radiology reports describe clinical findings and their attributes, offering an indirect signal of graded clinical relatedness.
- **Can relationships between reports help preserve fine-grained clinical differences, without severity annotations during pretraining?**


Rank-induced soft target
Text
Self-Similarity $S^{(l)}$
Row-wise Ranking
1.0 0.6 0.1 0.3
0.6 1.0 0.7 0.5
0.3 0.5 0.2 1.0
1 2 4 3
3 1 2 4
3 2 4 1
Rank-distance
Matrix $P^{(l)}$
Softmin
$\hat{Y}$
Cross-modal Similarity
$S^{(m)} = E^{(v)}E^{(l)^T}$
Predicted Distribution
$Q = softmax(S^{(m)}/\tau)$
Smoothed Global
Alignment Loss
$L_{global} = CE(\hat{Y}, Q)$


Rank-induced soft target & Modality-wise Similarity Discrepancy (MSD)
Text
Self-Similarity $S^{(l)}$
Row-wise Ranking
1.0 0.6 0.1 0.3
0.6 1.0 0.7 0.5
0.3 0.5 0.2 1.0
1 2 4 3
3 1 2 4
3 2 4 1
Rank-distance
Matrix $P^{(l)}$
Softmin
$\hat{Y}$
Visual
Self-Similarity $S^{(v)}$
Row-wise Ranking
1.0 0.1 0.6 0.2
0.1 1.0 0.0 0.5
0.2 0.5 0.3 1.0
1 4 2 3
3 1 4 2
4 2 3 1
Rank-distance
Matrix $P^{(v)}$
Modality-wise
Discrepancy
Adaptive
Temperature
τ_{MSD}
Cross-modal Similarity
$S^{(m)} = E^{(v)}E^{(l)^T}$
Temperature Scaling
$S^{(m)}/\tau_{MSD}$
Predicted Distribution
$Q = softmax(S^{(m)}/\tau_{MSD})$
Smoothed Global
Alignment Loss
$L_{global} = CE(\hat{Y}, Q)$

from MGCA (NeurIPS 2022)



Experiment

Zero-shot severity grading

Dataset	Edema Severity				CAD-Chest Atelectasis				CAD-Chest Pulmonary Edema				CAD-Chest Cardiomegaly			
Model	Acc	F1	MAE	QWK	Acc	F1	MAE	QWK	Acc	F1	MAE	QWK	Acc	F1	MAE	QWK
GLoRIA	12.18	9.96	1.593	0.104	31.53	13.89	1.223	0.079	36.38	27.87	0.854	0.380	31.45	19.93	1.210	0.166
MGCA	31.25	30.01	0.989	0.108	36.28	24.93	1.040	0.281	35.28	31.13	0.966	0.316	38.13	31.97	0.906	0.416
SAT	37.32	34.41	1.035	0.207	35.73	31.65	1.054	0.193	46.22	45.58	0.791	0.437	**44.45**	**41.97**	0.843	0.422
MedKLIP	30.78	18.69	1.249	0.020	21.13	11.17	1.644	-0.001	23.17	13.15	1.497	0.047	20.94	11.68	1.558	-0.020
CARZero	22.64	14.23	1.028	-0.090	21.90	15.81	1.320	-0.105	24.14	13.11	1.092	-0.023	26.45	20.25	1.244	-0.083
ASG	36.19	30.33	1.079	0.117	38.05	33.57	1.039	0.265	45.84	45.83	0.765	0.404	35.49	33.72	1.073	0.174
MAVL	25.81	14.86	0.963	-0.125	22.12	11.95	1.619	0.002	24.01	17.55	1.528	-0.096	27.55	19.21	1.062	0.060
RADZero	18.18	15.71	1.414	0.066	21.13	11.12	1.074	0.047	25.48	13.12	1.020	0.035	22.34	10.82	1.541	0.012
FinAl (ours)	**44.74**	**42.57**	**0.756**	**0.431**	**40.71**	**37.46**	**0.779**	**0.519**	**52.44**	**52.68**	**0.629**	**0.572**	42.84	41.72	**0.749**	**0.531**

- Lowest MAE and highest QWK across all four severity benchmarks.

Experiment

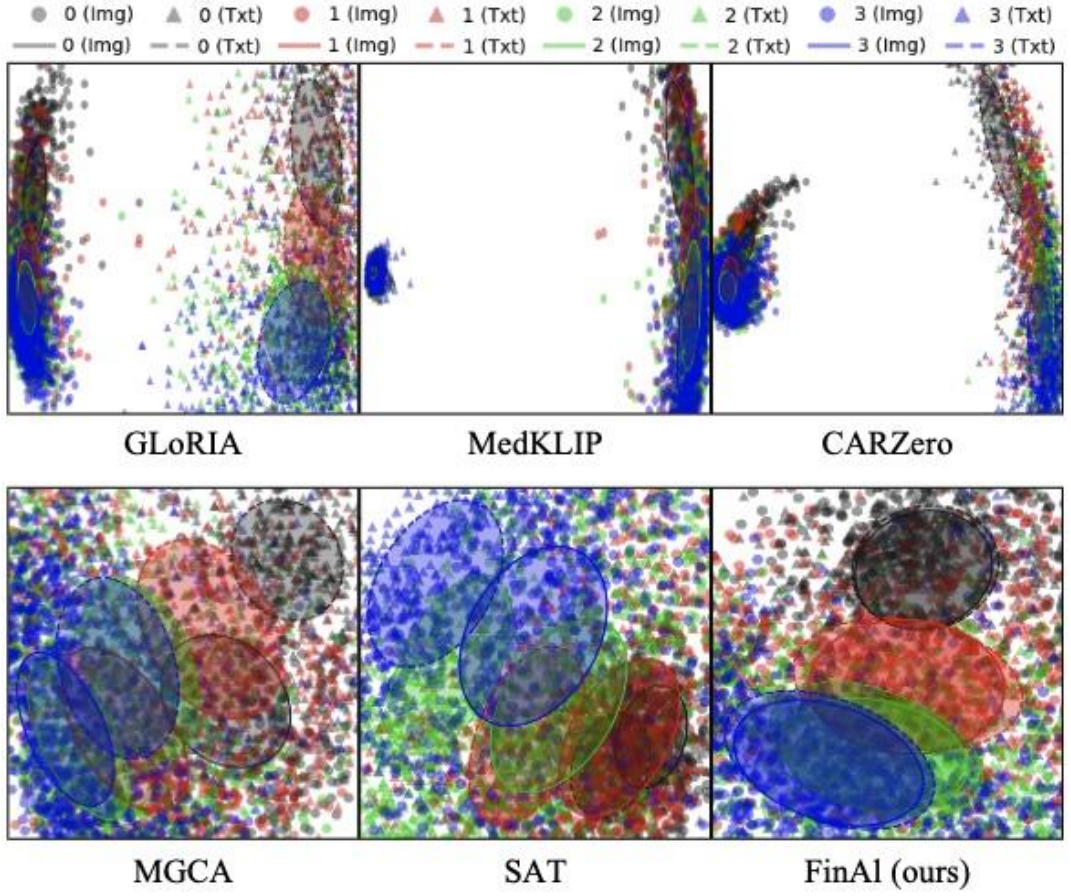
Severity-aware retrieval (Recall@1)

Dataset	Edema Severity			CAD Atelectasis			CAD Pulmonary Edema			CAD Cardiomegaly		
Model	I2I	I2T	T2I	I2I	I2T	T2I	I2I	I2T	T2I	I2I	I2T	T2I
GLoRIA	47.72	26.19	34.48	41.48	39.05	42.59	42.81	23.45	29.98	39.90	34.02	38.28
MGCA	42.15	33.73	41.08	40.60	47.23	53.10	47.60	42.81	50.67	43.86	47.17	49.45
SAT	39.54	44.92	42.51	41.92	58.41	56.31	48.23	52.73	55.05	42.47	55.77	54.08
MedKLIP	42.71	19.59	20.04	42.81	23.45	29.98	48.23	27.08	27.42	37.25	29.10	27.11
CARZero	51.04	26.90	10.96	42.92	23.12	32.63	49.03	27.71	25.57	42.98	24.25	23.14
ASG	52.00	47.29	49.83	39.49	57.08	60.18	50.00	52.94	54.84	45.48	56.43	53.42
MAVL	49.26	19.96	29.64	42.59	21.02	31.08	48.49	26.03	21.19	44.60	25.72	26.30
RADZero	40.12	37.10	28.72	29.76	27.10	27.88	37.72	25.44	21.70	33.14	18.81	27.85
FinAl(ours)	**54.18**	**52.38**	**51.90**	**45.13**	**71.02**	**69.03**	**50.13**	**63.16**	**58.66**	**45.63**	**66.94**	**59.52**

- Highest Recall@1 among the compared VLP models across all four severity benchmarks and all three retrieval directions.

Experiment

- PCA visualization of image and text representations on EdemaSeverity.



- FinAl shows more consistent overlap between image and text distributions of the same severity level.

Ablation Studies

- Loss component

Alignment		Edema Severity			
Fine	Local	Acc	Sens	Prec	F1
		43.78	45.04	42.75	41.29
	✓	40.36	44.26	42.92	39.16
✓		44.72	45.69	43.18	40.61
✓	✓	**44.74**	**46.55**	**43.91**	**42.57**

- Target construction

Target	Acc ↑	Macro-F1 ↑	MAE ↓	QWK ↑
Hard one-hot	40.36	39.16	0.85	0.33
Cosine-based soft target	37.49	27.59	0.98	0.23
Rank-induced (FinAl)	**44.74**	**42.57**	**0.76**	**0.43**

- Temperature scheduling

Scheduling	Acc	Sens	Prec	F1
Fixed ($\tau = 0.07$)	44.34	46.32	41.98	40.98
Learned	**45.15**	45.59	42.20	41.49
MSD (Ours)	44.74	**46.55**	**43.91**	**42.57**

Conclusion

- Radiology reports can supervise more than matched image-report correspondence.
- By converting report-similarity rankings into graded alignment targets, **FinAl** helps preserve fine-grained clinical information without severity annotations during pretraining.
- Severity grading and retrieval results show that **FinAl**'s pretrained representations better distinguish graded clinical states and preserve severity-aware relationships between cases.