Making the Most of Text Semantics to Improve Biomedical Vision-Language Processing

Benedikt Boecking*†, Naoto Usuyama*, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, Hoifung Poon, and Ozan Oktay‡

Presenter: Aylin Gunal

Problem Intro

1) Lack of large datasets of paired image, text data in medical domain
2) Prior work on VLP models ignores some of the potentials of refining the text modelling component

Proposed Solutions

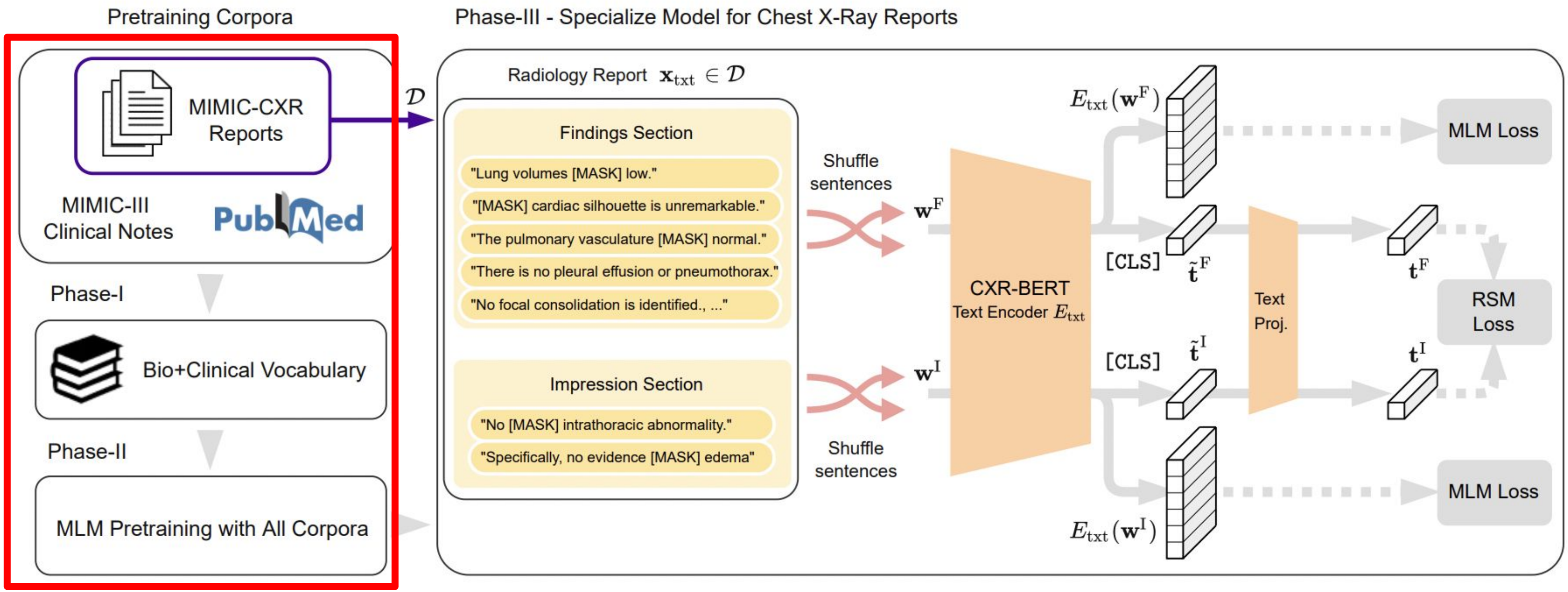
Overview: refine the text modeling component of the training process to improve overall performance of the VLP model

Specific contributions:

1) CXR-BERT
2) Novel approach to joint training in VLP
3) MS-CXR dataset

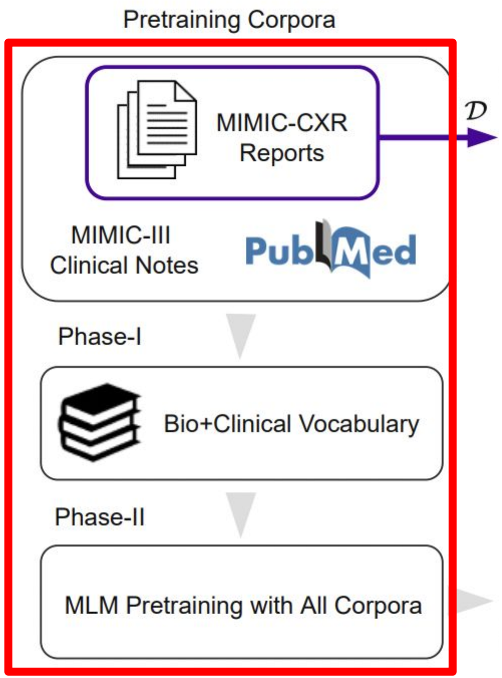
CXR-BERT

CXR-BERT



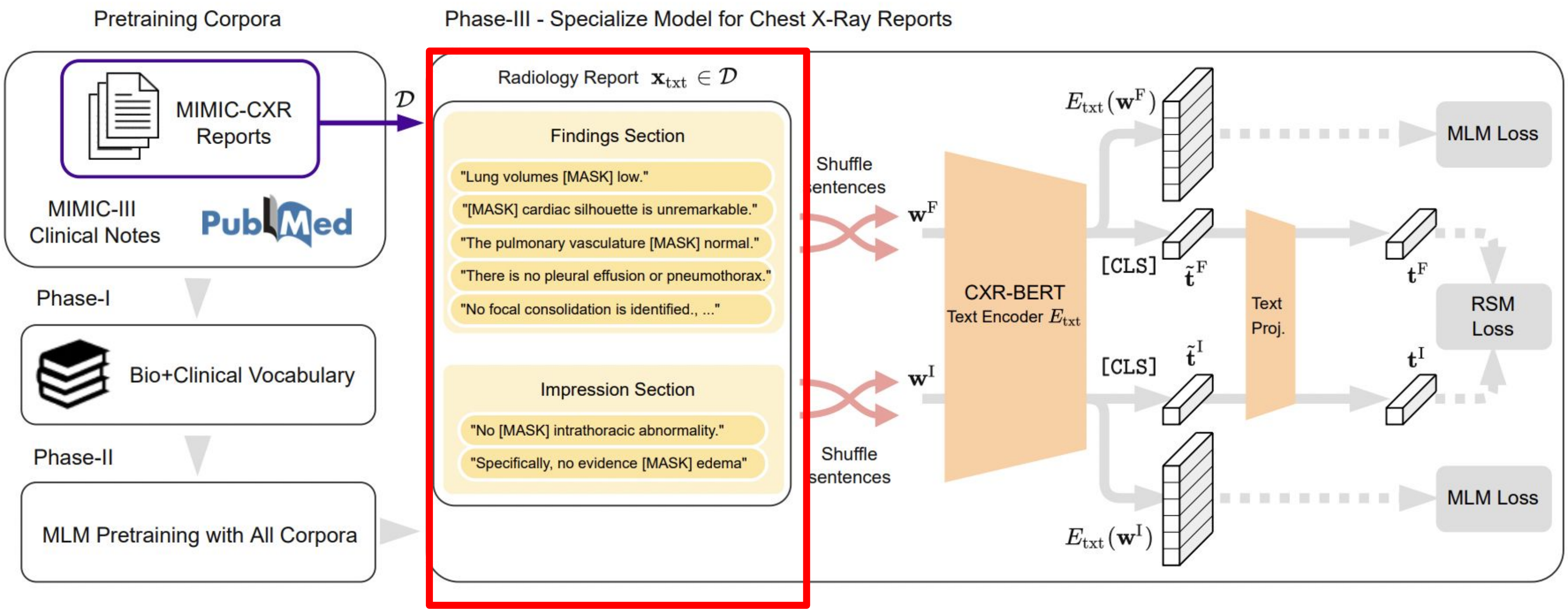
D = all documents in combined training corpus, comprised of articles from (1) PubMed abstracts; (2) MIMIC-III clinical notes; (3) MIMIC-CXR

CXR-BERT



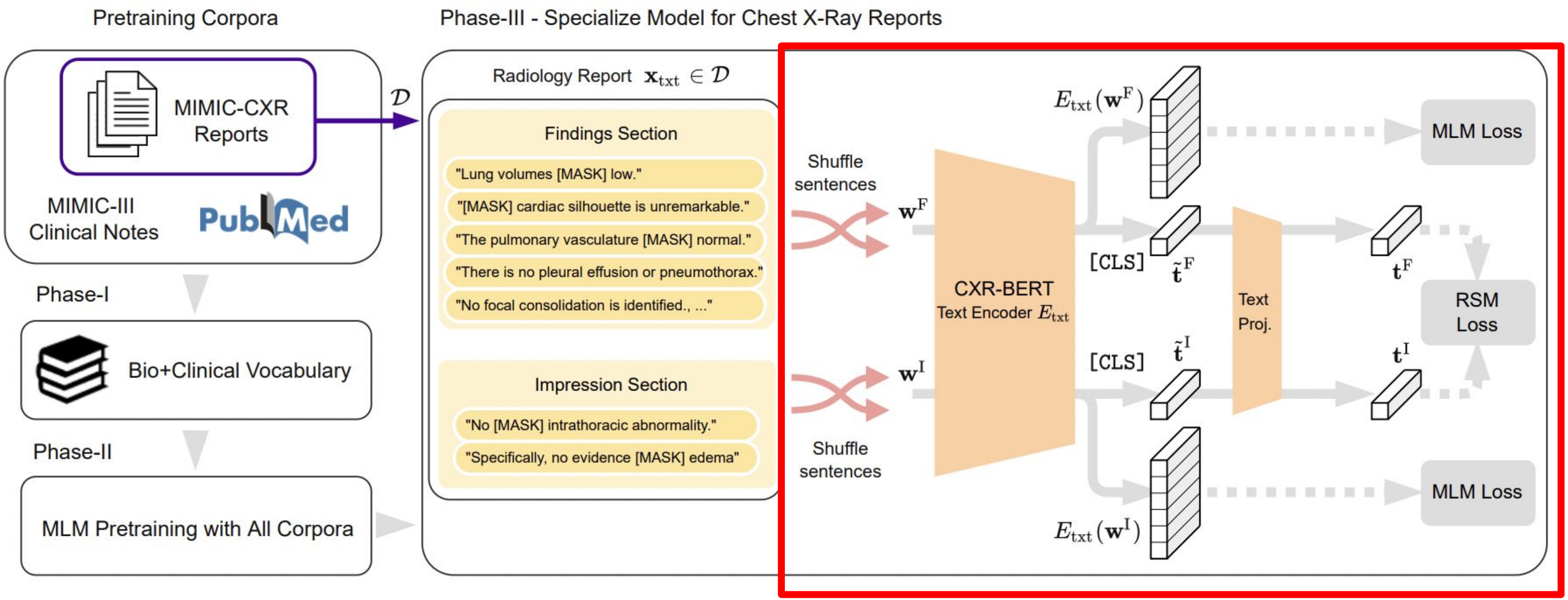
- D = all documents in combined training corpus, comprised of articles from
    - (1) PubMed abstracts
    - (2) MIMIC-III clinical notes
    - (3) MIMIC-CXR
- Phase-I
    - Create custom, domain-specific vocabulary
    - Motivation: avoid sub-word breakdowns common in the vocabulary in other similar, clinical language models
- Phase-II
    - Pre-train randomly initialized BERT model on MLM task on full training corpus
    - Use RoBERTa pre-training configurations
    - Motivation: build a general-domain specific language model (i.e. healthcare), which can then be ‘fine-tuned’ on CXR data (Phase III)

CXR-BERT



Findings = details on clinical findings
Impression = summarizing the clinical assessment

CXR-BERT



Authors introduce new pre-training task: radiology section matching (RSM), and the total loss is a combination of L_{RSM} and L_{MLM}.

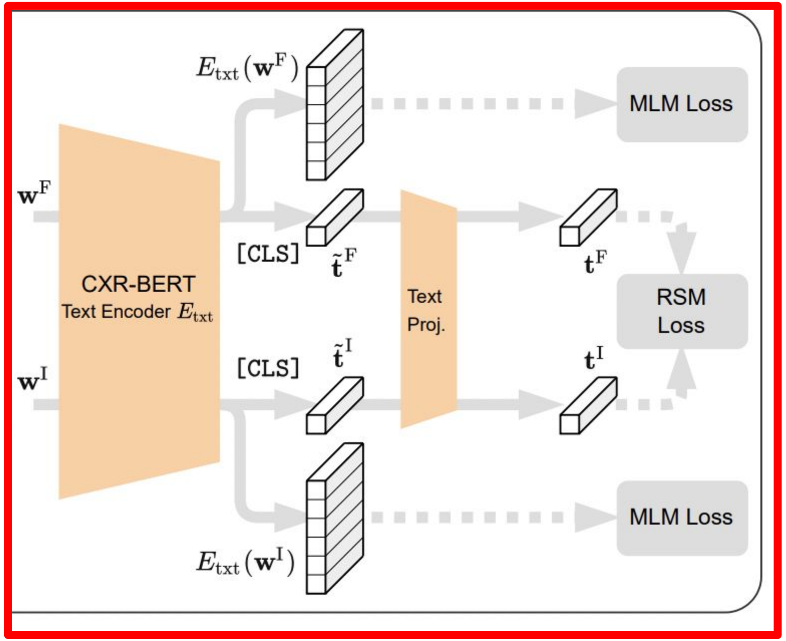
CXR-BERT

- w^F and w^I are sentences from Findings and Impressions sections, respectively
- RSM task: determine if w^F and w^I come from the same report (i.e. the same Findings, Impressions pair)
- CXR-BERT output
  - w^F and w^I embeddings are used for MLM loss

$$\mathcal{L}_{\mathrm{MLM}} = -\frac{1}{|\mathcal{B}|} \sum_{\mathbf{w} \in \mathcal{B}} \log p_\theta(\mathbf{w}_m \mid \mathbf{w}_{\setminus m})$$

  - [CLS] tokens of w^F and w^I are projected to a lower dimension using a two-layer perceptron

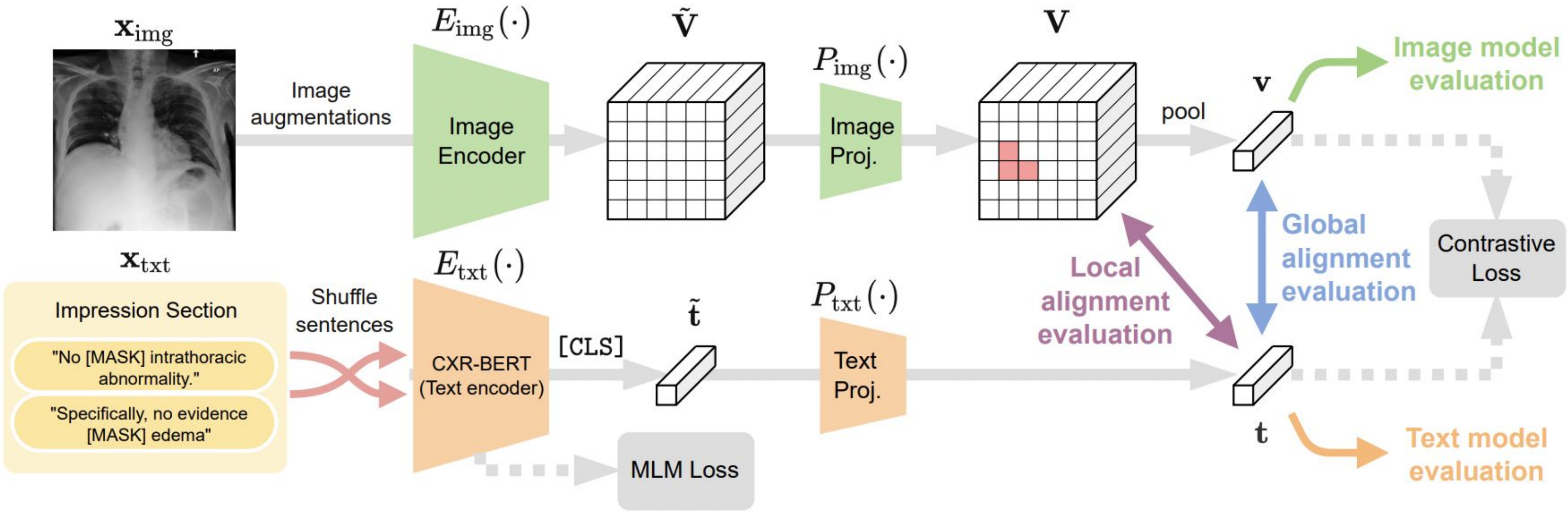
$$\mathcal{L}_{\mathrm{RSM}} = -\frac{1}{N} \sum_{i=1}^{N} \left(\log \frac{\exp(\mathbf{t}_i^{\mathrm{F}} \cdot \mathbf{t}_i^{\mathrm{I}} / \tau_1)}{\sum_{j=1}^{N} \exp(\mathbf{t}_i^{\mathrm{F}} \cdot \mathbf{t}_j^{\mathrm{I}} / \tau_1)} + \log \frac{\exp(\mathbf{t}_i^{\mathrm{I}} \cdot \mathbf{t}_i^{\mathrm{F}} / \tau_1)}{\sum_{j=1}^{N} \exp(\mathbf{t}_i^{\mathrm{I}} \cdot \mathbf{t}_j^{\mathrm{F}} / \tau_1)} \right)$$



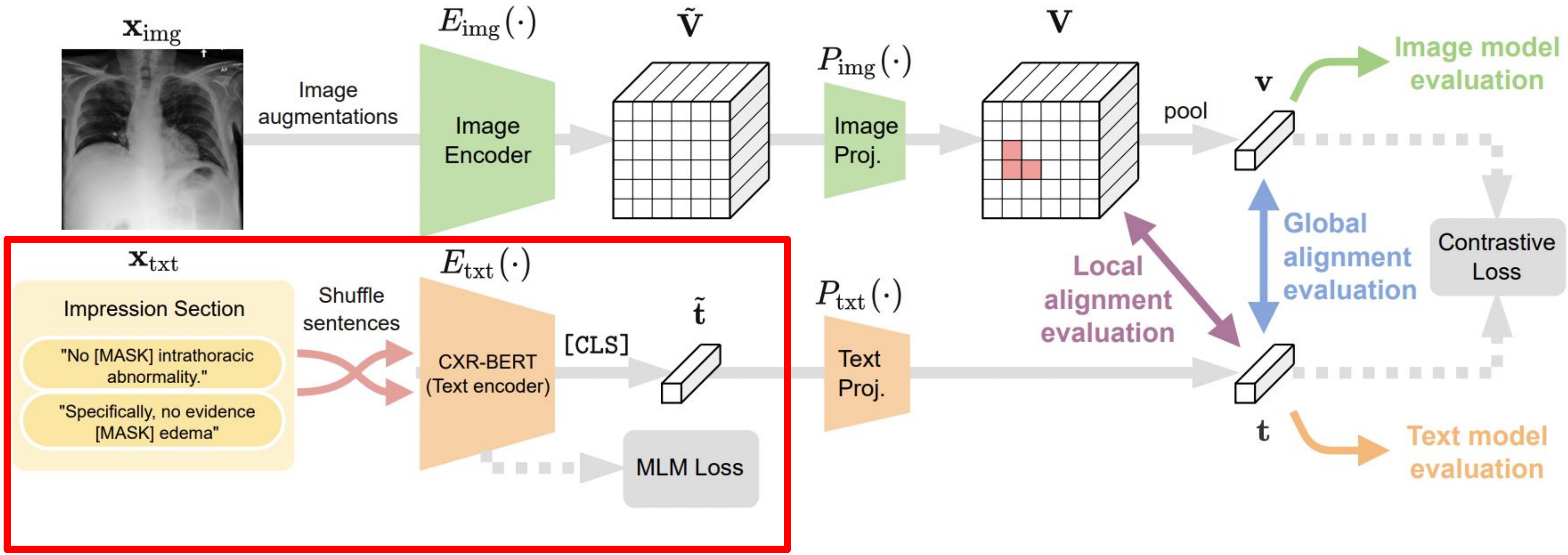
BioVil

BioVIL

Overview: a CNN image encoder and CXR-BERT text encoder are used to create representations that are then projected using 2-layer perceptrons, and those projections are then projected to a joint space

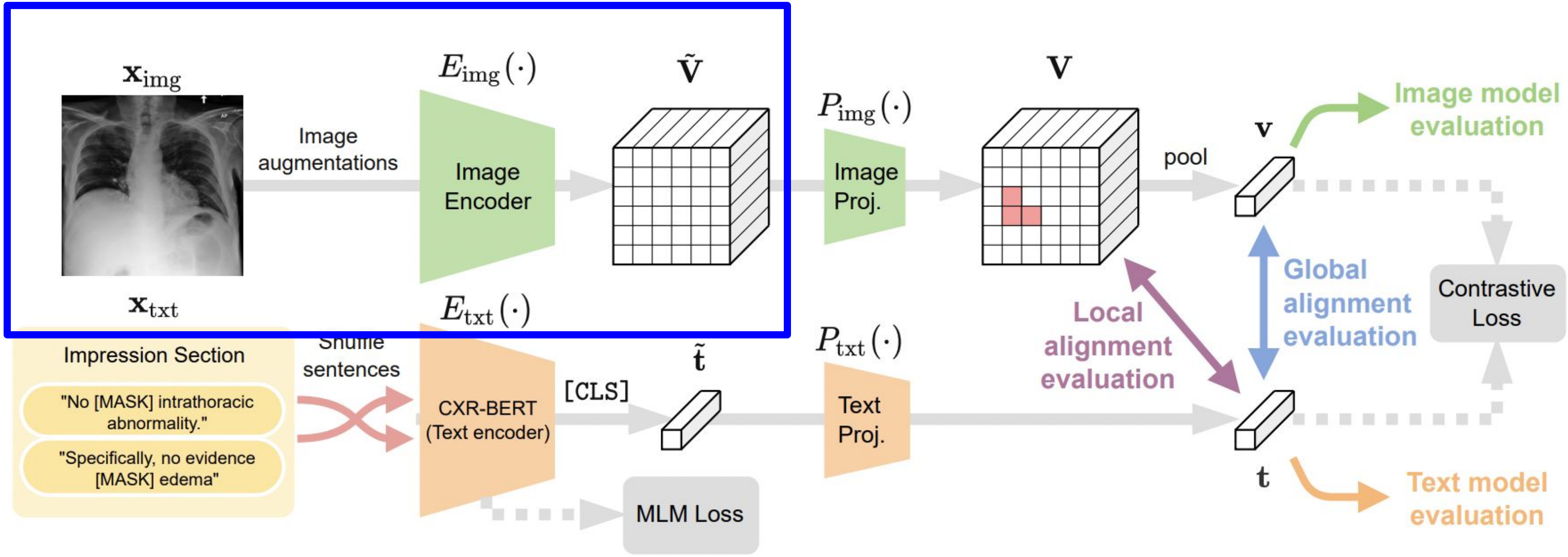


BioVIL



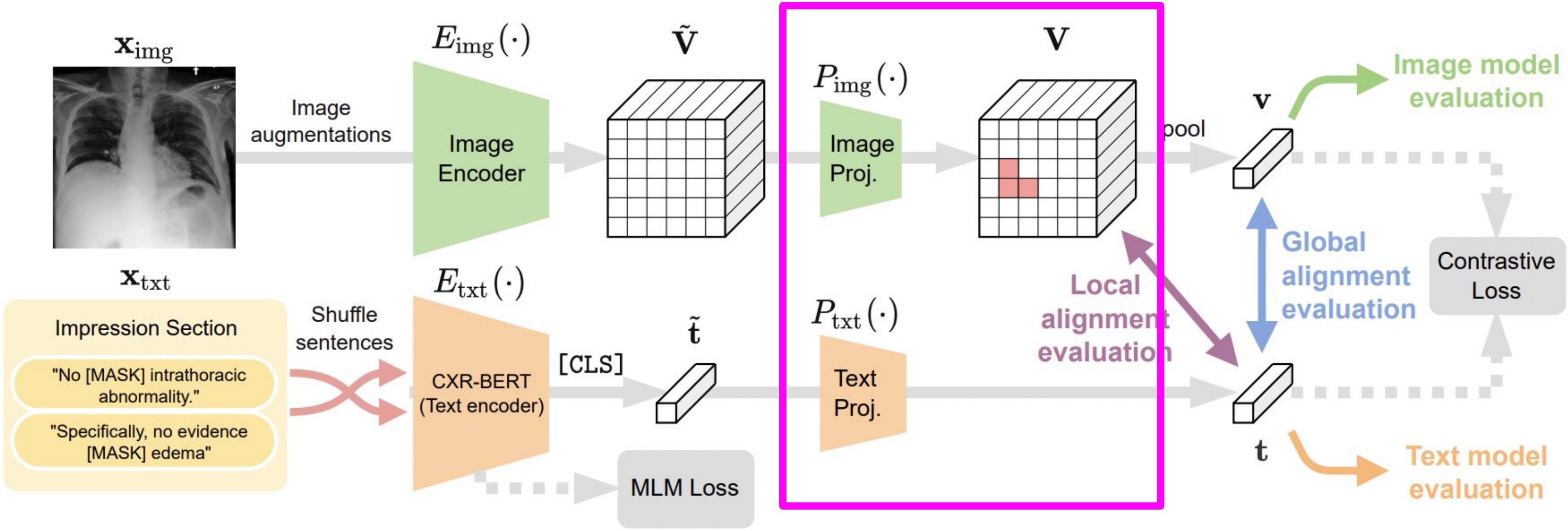
The text encoder, i.e. CXR-BERT as previously described (or theoretically some other domain-specific text encoder)

BioVIL



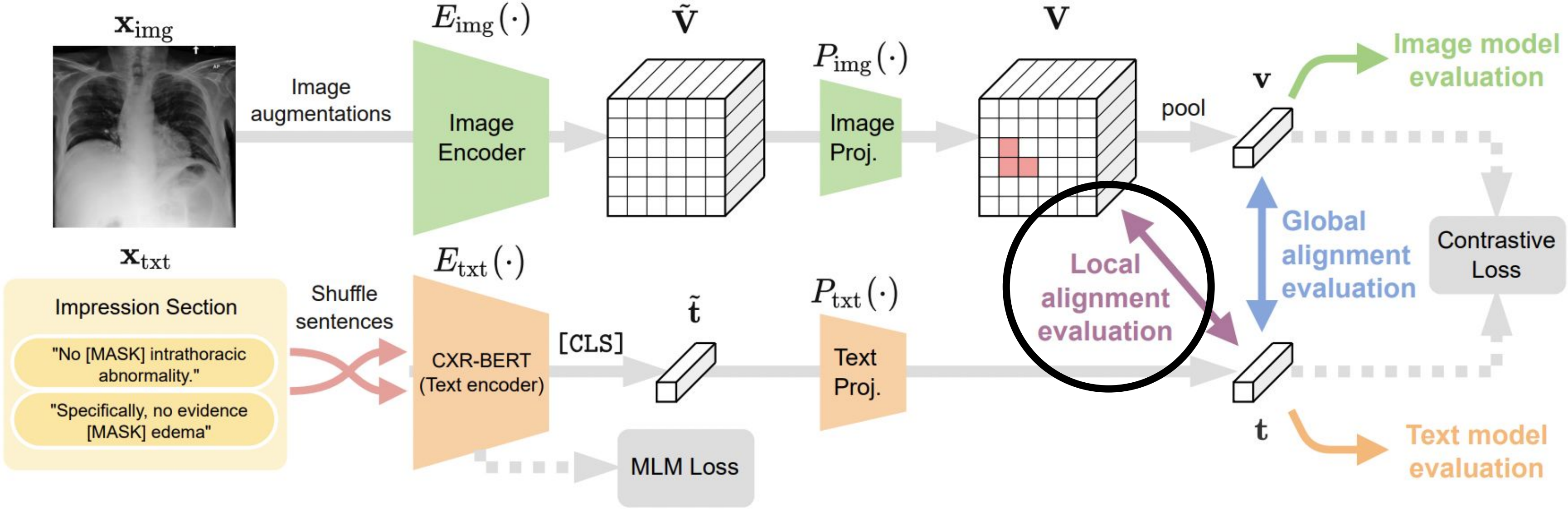
The image encoder, ResNet-50 pretrained on MIMIC-CXR images using SimCLR to output local image embeddings (V˜)

BioVIL



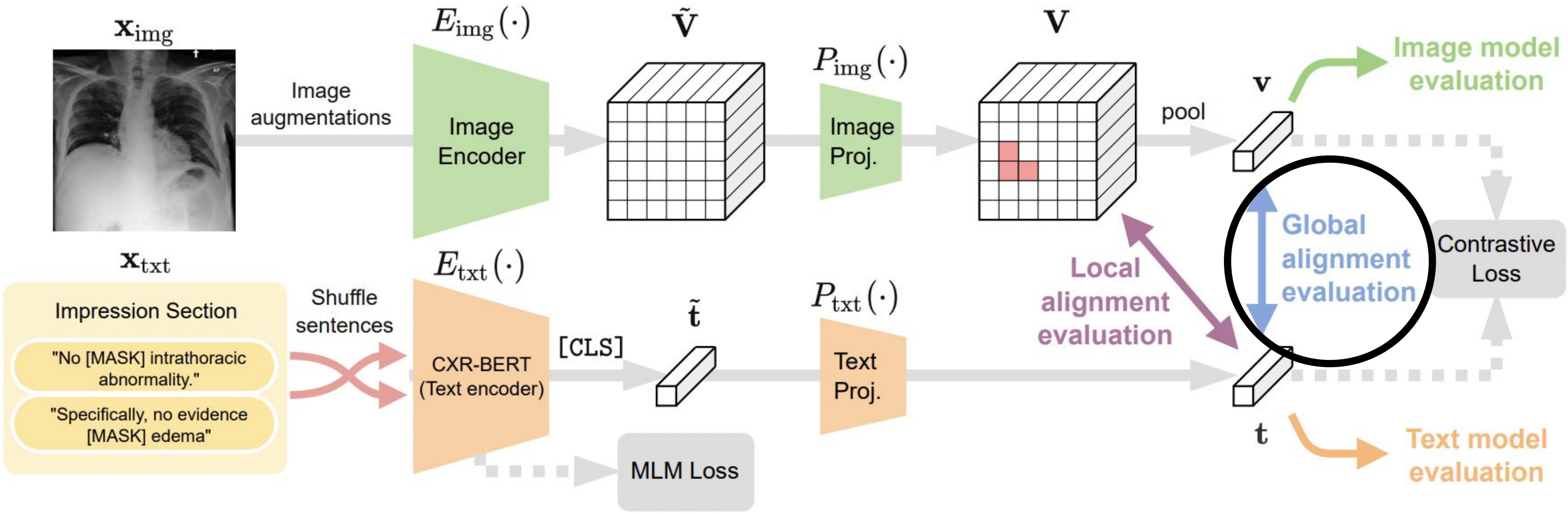
Both text and image get their own 2-layer perceptron projection models, which projects each embedding separately to a joint space of 128 dimensions, V

BioVIL



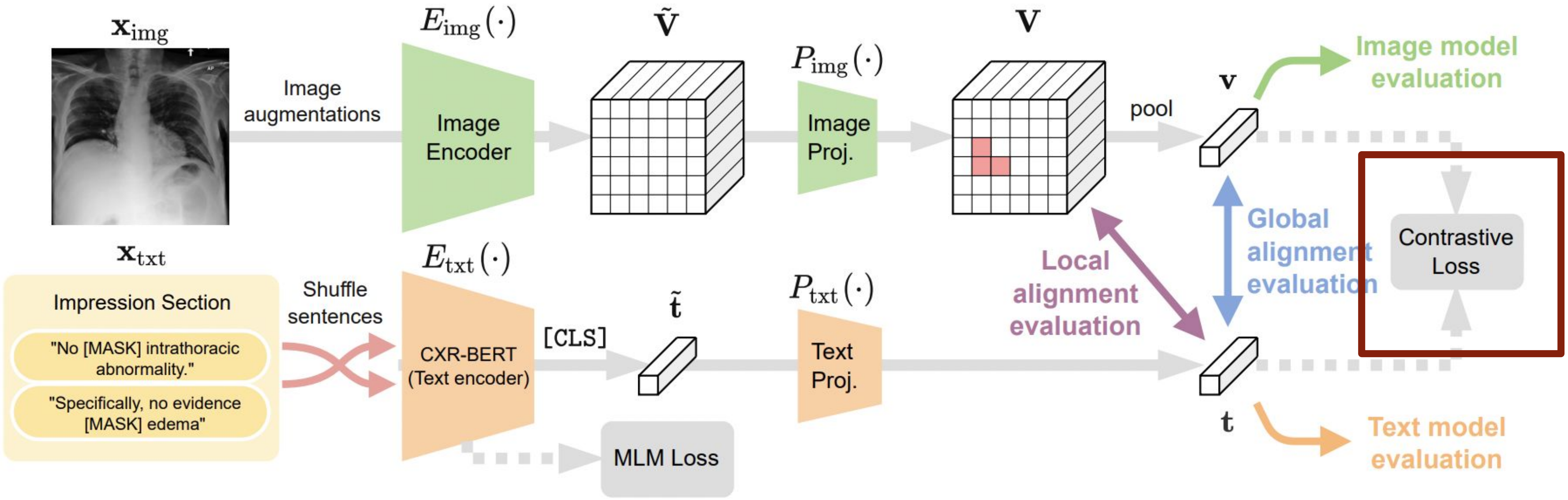
$\mathbf{x}_{\text{img}}$
Image augmentations
$E_{\text{img}}(\cdot)$
Image Encoder
$\tilde{\mathbf{V}}$
$P_{\text{img}}(\cdot)$
Image Proj.
$\mathbf{V}$
pool
$\mathbf{v}$
Image model evaluation
Global alignment evaluation
Contrastive Loss
Local alignment evaluation
$\mathbf{x}_{\text{txt}}$
Impression Section
"No [MASK] intrathoracic abnormality."
"Specifically, no evidence [MASK] edema"
Shuffle sentences
$E_{\text{txt}}(\cdot)$
CXR-BERT (Text encoder)
[CLS]
$\tilde{\mathbf{t}}$
$P_{\text{txt}}(\cdot)$
Text Proj.
$\mathbf{t}$
Text model evaluation
MLM Loss

BioVIL



$\mathbf{x}_{\text{img}}$
Image augmentations
$E_{\text{img}}(\cdot)$
Image Encoder
$\tilde{\mathbf{V}}$
$P_{\text{img}}(\cdot)$
Image Proj.
$\mathbf{V}$
pool
$\mathbf{v}$
Image model evaluation
Global alignment evaluation
Contrastive Loss
$\mathbf{x}_{\text{txt}}$
Impression Section
"No [MASK] intrathoracic abnormality."
"Specifically, no evidence [MASK] edema"
Shuffle sentences
$E_{\text{txt}}(\cdot)$
CXR-BERT (Text encoder)
[CLS]
$\tilde{\mathbf{t}}$
$P_{\text{txt}}(\cdot)$
Text Proj.
Local alignment evaluation
$\mathbf{t}$
Text model evaluation
MLM Loss

BioVIL



The final loss term utilizes both information from the joint training process and the global alignment measure

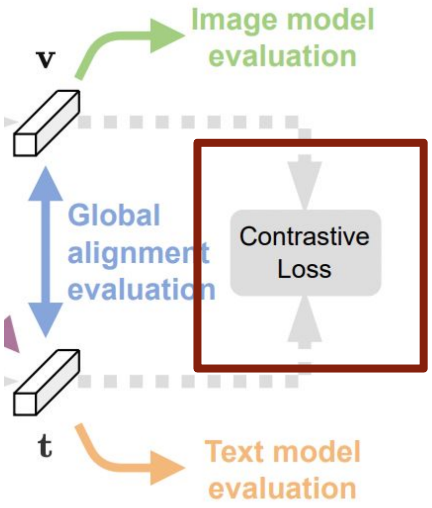
BioVIL

- Total loss is comprised of two terms
  - L_{MLM} → carries over from CXR-BERT

$$\mathcal{L}_{\text{MLM}} = -\frac{1}{|\mathcal{B}|} \sum_{\mathbf{w} \in \mathcal{B}} \log p_\theta(\mathbf{w}_m \mid \mathbf{w}_{\setminus m})$$

  - L_{LGA}

$$\mathcal{L}_{\text{GA}} = -\frac{1}{N} \sum_{i=1}^{N} \left(\log \frac{\exp(\mathbf{v}_i \cdot \mathbf{t}_i^{\text{I}} / \tau_2)}{\sum_{j=1}^{N} \exp(\mathbf{v}_i \cdot \mathbf{t}_j^{\text{I}} / \tau_2)} + \log \frac{\exp(\mathbf{t}_i^{\text{I}} \cdot \mathbf{v}_i / \tau_2)}{\sum_{j=1}^{N} \exp(\mathbf{t}_i^{\text{I}} \cdot \mathbf{v}_j / \tau_2)} \right). \quad (2)$$



BioVIL

Joint loss: $$\mathcal{L}_{\text{joint}} = \lambda_{\text{GA}}\mathcal{L}_{\text{GA}} + \mathcal{L}_{\text{MLM}}$$

Parameter values determined with grid search

MS-CXR

MS-CXR: a CXR Phrase Grounding Benchmark Dataset

- Collection and annotation: 2 board-certified radiologists annotate samples from the MIMIC-CXR dataset
- Format: Composed of pairs of image bounding box labels and radiology text descriptions
    - Descriptions are not just brief captions; in-depth descriptions

Findings	# of annotation pairs	# of subjects	Gender - F (%)	Avg Age (std)
Atelectasis	61	61	28 (45.90%)	64.52 (15.95)
Cardiomegaly	333	282	135 (47.87%)	68.10 (14.81)
Consolidation	117	109	40 (36.70%)	60.08 (17.67)
Edema	46	42	18 (42.86%)	68.79 (14.04)
Lung opacity	81	81	33 (40.24%)	62.07 (17.20)
Pleural effusion	96	95	41 (43.16%)	66.36 (15.29)
Pneumonia	182	146	65 (44.52%)	64.32 (17.17)
Pneumothorax	237	151	66 (43.71%)	60.71 (18.04)
Total	1153	851	382 (44.89%)	64.37 (16.61)
Background (all MIMIC-CXR)	-	65379	34134.0 (52.39%)	56.85 (19.47)

Experiments — CXR-BERT alone

	RadNLI accuracy (MedNLI transfer)	Mask prediction accuracy	Avg. # of tokens after tokenization	Vocabulary size
RadNLI baseline [54]	53.30	-	-	-
ClinicalBERT	47.67	39.84	78.98 (+38.15%)	28,996
PubMedBERT	57.71	35.24	63.55 (+11.16%)	28,895
CXR-BERT (after Phase-III)	60.46	77.72	58.07 (+1.59%)	30,522
CXR-BERT (after Phase-III + Joint Training)	65.21	81.58	58.07 (+1.59%)	30,522

Experiments — BioVIL Joint Set-up

Method	Type	Text model	Loss	% of labels	Acc.	F1	AUROC
SimCLR [6]	Image only	-	Global	1%	0.545	0.522	0.701
				10%	0.760	0.639	0.802
				100%	0.788	0.675	0.849
GLoRIA [31]	Joint	ClinicalBERT	Global & local	Zero-shot	0.70	0.58	-
				1%	0.72	0.63	0.861
				10%	0.78	0.63	0.880
				100%	0.79	0.65	0.886
Baseline	Joint	ClinicalBERT	Global	Zero-shot	0.719	0.614	0.812
BioViL	Joint	CXR-BERT	Global	Zero-shot	0.732	0.665	0.831
				1%	0.805	0.723	0.881
				10%	0.812	0.727	0.884
				100%	$\boxed{0.822}$	$\boxed{0.733}$	$\boxed{0.891}$

Experiments — BioVIL Zero-shot and linear probing

Method	% of Labels	Supervision	IoU	Dice	CNR
LoVT [56]	100%	Lin. prob.	-	0.518	-
ConVIRT [85]	-	Zero-shot	0.228	0.348	0.849
GLoRIA [31]	-	Zero-shot	0.245	0.366	1.052
BioViL	-	Zero-shot	0.355	0.496	1.477
SimCLR [6]	5%	Lin. prob.	0.382	0.525	1.722
SimCLR [6]	100%	Lin. prob.	0.427	0.570	1.922
BioViL	5%	Lin. prob.	0.446	0.592	2.077
BioViL	100%	Lin. prob.	0.469	0.614	2.178

Quiz [with answers]

1) Why is random sentence shuffling feasible and successful in the CXR-BERT text augmentation process?
   a) Answer: the order of the clinical notes for these particular data (i.e. CXR impressions and findings sections) generally doesn't matter for sentences in the findings and the assessments
2) How can the RSM loss formula be interpreted?
   a) Answer: rewarding for classifying items in the same pair as so (true positive), while simultaneously penalizing classification of items not in the same pair as items in the same pair (false positive)

$$\mathcal{L}_{\mathrm{RSM}} = -\frac{1}{N}\sum_{i=1}^{N}\left(\log\frac{\exp(\mathbf{t}_i^{\mathrm{F}}\cdot\mathbf{t}_i^{\mathrm{I}}/\tau_1)}{\sum_{j=1}^{N}\exp(\mathbf{t}_i^{\mathrm{F}}\cdot\mathbf{t}_j^{\mathrm{I}}/\tau_1)} + \log\frac{\exp(\mathbf{t}_i^{\mathrm{I}}\cdot\mathbf{t}_i^{\mathrm{F}}/\tau_1)}{\sum_{j=1}^{N}\exp(\mathbf{t}_i^{\mathrm{I}}\cdot\mathbf{t}_j^{\mathrm{F}}/\tau_1)}\right)$$