The Role of Local Alignment and Uniformity in Image-Text Contrastive Learning on Medical Images

Philip Müller
Technical University of Munich
philip.j.mueller@tum.de

Georgios Kaissis
Technical University of Munich
Helmholtz Zentrum Munich

Daniel Rueckert
Technical University of Munich
Imperial College London

Abstract

Image-text contrastive learning has proven effective for pretraining medical image models. When targeting localized downstream tasks like semantic segmentation or object detection, additional local contrastive losses that align image regions with sentences have shown promising results. We study how local contrastive losses are related to global (per-sample) contrastive losses and which effects they have on localized medical downstream tasks. Based on a theoretical comparison, we propose to remove some components of local losses and replace others by a novel distribution prior which enforces uniformity of representations within each sample. We empirically study this approach on chest X-ray tasks and find it to be very effective, outperforming methods without local losses on 12 of 18 tasks.

1 Introduction

Image-text contrastive learning [Radford et al., 2021, Jia et al., 2021] has been well established as a pretraining method for image models recently. By utilizing companion text like radiological reports or captions, it can improve the downstream performance of image models on tasks like image classification. Such image-text methods have also proven effective on medical images like chest X-rays [Zhang et al., 2020, Müller et al., 2022a]. Typically, such methods use (global) contrastive losses to align per-sample representation of images and the related text. However, when targeting localized downstream tasks like semantic segmentation or object detection, it has proven beneficial to also introduce local contrastive losses to align image regions (e.g. patches) with sentences [Müller et al., 2022b, Liao et al., 2021].

In this work we study how local contrastive losses are related to global contrastive losses and which effects they have on localized downstream tasks. Following Wang and Isola [2020], we decompose the global and local losses into alignment components (pulling representations close to each other) and distribution priors (pushing representations away from each other). We found that the alignment components of global and local losses are related and assume that they have similar effects, while the distribution priors of global and local losses have complementary effects. We empirically study these findings on localized downstream tasks on chest X-rays and therefore propose a pre-training method consisting of a global contrastive loss and local uniformity regularizers (i.e. distribution priors) but without local alignment components. Our results show that this method typically performs well on tasks where local contrastive methods outperform global contrastive methods, thus proving our assumptions and indicating the relevance of local uniformity for localized downstream tasks. For simplicity, we focus our study on the LoVT method [Müller et al., 2022b], but argue that our findings are also relevant for related approaches (see Appendix A).

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

2 Analysis of the Relation between Global and Local Contrastive Losses

We now theoretically study the relationship of the global and local contrastive losses of LoVT [Müller et al., 2022b]. For a detailed derivation we refer to Appendix C. Like many image-text contrastive works, LoVT uses two independent encoders (one for images and one for text, i.e. radiological reports) to compute global (i.e. per-image/per-report) and local (i.e. patch/sentence) representations. The global image and text representations (denoted by $\bar{\boldsymbol{z}}_i^{\mathcal{I}}$ and $\bar{\boldsymbol{z}}_i^{\mathcal{R}}$) are then aligned using a NTXent-based [Chen et al., 2020] contrastive loss, denoted by $\mathcal{L}_{\text{global}}$. The local representations are first transferred to the other modality using an attention model, such that for each patch representation $\boldsymbol{z}_{i,k}^{\mathcal{I}}$ (of patch k from the image model) we have a representation $\boldsymbol{z}_{i,k}^{\mathcal{R}\to\mathcal{I}}$ with information from the report and vice-versa for each sentence representation $\boldsymbol{z}_{i,m}^{\mathcal{R}}$ (of sentence m) we have a representation $\boldsymbol{z}_{i,m}^{\mathcal{I}\to\mathcal{R}}$ with information from the image. The representation pairs $(\boldsymbol{z}_{i,k}^{\mathcal{I}}, \boldsymbol{z}_{i,k}^{\mathcal{R}\to\mathcal{I}})$ and $(\boldsymbol{z}_{i,m}^{\mathcal{R}}, \boldsymbol{z}_{i,m}^{\mathcal{I}\to\mathcal{R}})$ are then aligned using the local NTXent-based losses $\mathcal{L}_{\text{local-image}}$ and $\mathcal{L}_{\text{local-report}}$, respectively. Following Wang and Isola [2020], each of the three contrastive losses can be decomposed into an *alignment* component, maximizing the similarity of related representations (i.e. pulling them close to each other), and a *distribution prior* component, imposing a distribution on the representations to prevent them from collapsing to a constant (i.e. pushing non-related representations away form each other). We can therefore decompose the three losses into three alignment components ($\mathcal{L}_{\text{global-align}}$ for global, $\mathcal{L}_{\text{local-align}}^{\mathcal{I}}$ for local image, and $\mathcal{L}_{\text{local-align}}^{\mathcal{R}}$ for local text alignment) and three distribution priors ($\mathcal{L}_{\text{global-dist}}$ as global, $\mathcal{L}_{\text{local-dist}}^{\mathcal{I}}$ as local image, and $\mathcal{L}_{\text{local-dist}}^{\mathcal{R}}$ as local report prior).

2.1 Alignment Components

In order to study the relations between the alignment components $\mathcal{L}_{\text{global-align}}$, $\mathcal{L}_{\text{local-align}}^{\mathcal{I}}$, and $\mathcal{L}_{\text{local-align}}^{\mathcal{R}}$, we make some simplifying assumptions: i) We assume that the representations $\bar{\boldsymbol{z}}_i^{\mathcal{I}}$ and $\boldsymbol{z}_{i,m}^{\mathcal{I}\to\mathcal{R}}$ are computed as weighted sums from region representations $\boldsymbol{z}_{i,k}^{\mathcal{I}}$, and similarly $\bar{\boldsymbol{z}}_i^{\mathcal{R}}$ and $\boldsymbol{z}_{i,k}^{\mathcal{R}\to\mathcal{I}}$ as weighted sums from sentence representations $\boldsymbol{z}_{i,m}^{\mathcal{R}}$, and ii) we ignore the normalization terms of the cosine similarity, i.e. we treat all cosine similarities as dot products, denoted by $\text{dot}(\cdot,\cdot)$. Using these assumptions, and given the batch size N, with K patches per image, and M_i sentences in sample i, each of the three alignment components can be written as a weighted sum (with weights $\xi_{i,k,m}$) of dot products between region representations $\boldsymbol{z}_{i,k}^{\mathcal{I}}$ and sentence representations $\boldsymbol{z}_{i,m}^{\mathcal{R}}$:

$$\mathcal{L}_{\text{align}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{m=1}^{M_i}\xi_{i,k,m}\text{dot}\left(\boldsymbol{z}_{i,k}^{\mathcal{I}}, \boldsymbol{z}_{i,m}^{\mathcal{R}}\right) , \tag{1}$$

where the exact form of weights $\xi_{i,k,m}$ depends on the specific loss function (see Appendix C.4 for details). Therefore, the only difference between the three alignment components is the definition of $\xi_{i,k,m}$. Assuming there are better and worse aligned pairs of local representations in each sample (i.e. pair of image and report), the main difference between those losses is how the best aligned pairs are weighted against the worse aligned pairs. This means that in the special case where the local representations of each modality are constant within each sample, they are identical. One major difference between global and local alignment components is, that for local alignments $\mathcal{L}_{\text{local-align}}^{\mathcal{I}}$ and $\mathcal{L}_{\text{local-align}}^{\mathcal{R}}$, the $\xi_{i,k,m}$ are not separable into components containing only information from a single modality (image or report). The reason for this is that the attention weights used to compute $\boldsymbol{z}_{i,k}^{\mathcal{R}\to\mathcal{I}}$ and $\boldsymbol{z}_{i,m}^{\mathcal{I}\to\mathcal{R}}$ are not separable in that way, as they are computed from both, image and report, representations. They therefore allow for pairwise interactions between both modalities, which are not possible with the global alignment component $\mathcal{L}_{\text{global-align}}$ (as $\xi_{i,k,m}$ is separable in that case). Summarizing our findings, we identified the following main differences between the global alignment and the local alignment components:

- Local alignment allows for more complex pairwise interactions between local representations, not restricted by the separability of the attention weights.
- Local and global alignment differ in how they weight well aligned pairs of local representations compared to less aligned pairs. Local alignment can incorporate pairwise distances between $\boldsymbol{z}_{i,k}^{\mathcal{I}}$ and $\boldsymbol{z}_{i,m}^{\mathcal{R}}$ and may thus focus on the best aligned pairs.
- Local and global alignment losses differ in where normalization (of the cosine similarity) happens, i.e. before or after the summation over local representations.

- Using global and local alignment together allows for decoupling of local from global representations by using independent projection heads.

Despite these differences, we argue that local and global alignment losses both enforce well aligned local as well as global representations. In Sec. 3 we therefore empirically study removing the local while keeping only the global alignment components.

2.2 Distribution Prior Components and Uniformity

Studying the distribution prior components $\mathcal{L}_{\text{global-dist}}$, $\mathcal{L}^{\mathcal{I}}_{\text{local-dist}}$, and $\mathcal{L}^{\mathcal{R}}_{\text{local-dist}}$, we found that they all follow a similar form: They all minimize the (weighted) average of `logsumexp`-aggregations over pairwise cosine-similarities between weighted sums of the local representations $\boldsymbol{z}^{\mathcal{I}}_{i,k}$ and $\boldsymbol{z}^{\mathcal{R}}_{i,m}$ (see Appendix C.5). However, while in the global distribution prior $\mathcal{L}_{\text{global-dist}}$ the `logsumexp`-aggregation is done over the samples in the batch, in the local distribution priors $\mathcal{L}^{\mathcal{I}}_{\text{local-dist}}$ and $\mathcal{L}^{\mathcal{R}}_{\text{local-dist}}$ the `logsumexp`-aggregation is done over the local representations of each sample (i.e. we sum over the regions or sentences per sample). This means that while the $\mathcal{L}_{\text{global-dist}}$ loss imposes representations to be roughly uniformly distributed over the whole dataset (i.e. pushing representations from different samples away from each other), $\mathcal{L}^{\mathcal{I}}_{\text{local-dist}}$ and $\mathcal{L}^{\mathcal{R}}_{\text{local-dist}}$ impose uniform distributions of (local) representations over each sample (i.e. pushing representations within each sample away from other representations from the same sample).

We argue that the contrast between local representations (including patch representations), which is enforced by the local distributions priors, is essential for the success of pretraining methods on localized downstream tasks. Therefore, instead of removing these local distribution priors, we propose to replace them by distribution priors that impose a per-sample uniform distribution on each modality (image and text) independently. Following Wang and Isola [2020], we use a loss based on pairwise Gaussian potentials but adapt it to impose uniform distributions within each sample instead of imposing them over the whole dataset. This *Gaussian uniformity (uni-gauss)* loss is then applied to the image and report modality independently. For images it is defined as

$$\mathcal{L}^{\mathcal{I}}_{\text{uni-gauss}} = \frac{1}{N}\sum_{i=1}^{N}\log\frac{1}{K^2}\sum_{k=1}^{K}\sum_{k'=1}^{K}\exp\left(-\frac{\cos(\boldsymbol{z}^{\mathcal{I}}_{i,k},\boldsymbol{z}^{\mathcal{I}}_{i,k'})}{\tau'}\right), \tag{2}$$

while for reports it is defined as

$$\mathcal{L}^{\mathcal{R}}_{\text{uni-gauss}} = \frac{1}{N}\sum_{i=1}^{N}\log\frac{1}{M_i^2}\sum_{m=1}^{M_i}\sum_{m'=1}^{M_i}\exp\left(-\frac{\cos(\boldsymbol{z}^{\mathcal{R}}_{i,m},\boldsymbol{z}^{\mathcal{R}}_{i,m'})}{\tau'}\right). \tag{3}$$

We will empirically study the effect of this loss in the next section.

3 Empirical Study

3.1 Experimental Setup

Following the insights from Sec. 2, we empirically study the effect of removing local alignment components and replacing local distribution components of LoVT's loss function, resulting in:

$$\mathcal{L}_{\text{LoVT-uni-gauss}} = \gamma\cdot\mathcal{L}_{\text{global}} + \eta\cdot\left(\mathcal{L}^{\mathcal{I}}_{\text{uni-gauss}} + \mathcal{L}^{\mathcal{R}}_{\text{uni-gauss}}\right), \tag{4}$$

where γ (we used the same value as LoVT) and η (determined by hyperparameter tuning) are loss coefficients. Apart from the changes to the loss function, we follow the same framework and training procedure as used in LoVT. Note however that we share the projection heads for local and global representations (per modality). We pretrain on 30% of MIMIC-CXRv2 [Johnson et al., 2019a,b,c] and then evaluate our models trained on the same evaluation framework [Müller et al., 2022a] as used in LoVT, which consists of 18 localized medical tasks (semantic segmentation and object detection) on five public chest X-ray datasets. We compare the results against methods only having global losses, i.e. CLIP [Radford et al., 2021], ConVIRT [Zhang et al., 2020], and LoVT with the local contrastive losses removed (LoVT /wo local), and against the unmodified LoVT. Note that we experimented with different temperatures τ' and also studied a cross-entropy-based variant of the proposed uniformity loss, which we call *uni-xent* (see Appendix C.5.2), but found the Gaussian uniformity loss to be more effective. We refer to Appendix D for detailed ablation studies.

Table 1: Correlation of downstream results of our uniformity-based modification of LoVT and the unmodified LoVT. We consider (in the rows) the number of tasks where uniformity-based models i) outperform CLIP, ii) outperform LoVT, iii) outperform LoVT /wo local, and iv) outperform all studied methods, and (in the columns) on how many of them a) LoVT outperforms CLIP (↑) or vice-versa (↓) and b) LoVT /wo local performs worse (↓) or better (↑) than LoVT with local losses.

	LoVT vs. CLIP		LoVT /wo local		Total	Total
	LoVT ↑	CLIP ↑	↓	↑	$> 95\%$ conf.	
Uni $>$ CLIP	**10 (4)**	4 (3)	**10 (5)**	4 (2)	10	of 14
Uni $>$ LoVT	**5** (1)	4 **(3)**	**6** (1)	3 **(2)**	6	of 9
Uni $>$ LoVT /wo local	**10 (4)**	5 **(4)**	**12 (7)**	3 (2)	12	of 15
Uni is best	**6 (2)**	3 **(2)**	**7 (2)**	2 (1)	4	of 9
Total	of **12 (5)**	of 6 (4)	of **13 (7)**	of 5 (2)	of 18	of 18

Note: The numbers in paranthesis always specify the number of tasks where the column is true for more than the 95%-confidence interval (of the better model). Bold numbers indicate that the cell is the best in its row and block. Underlined numbers indicate that the cells relative number of tasks (normalized by the column total) is larger than of the other cells in the same row and block. Cells containing a majority of tasks (more than half of the total tasks) in their column are marked with blue background.

3.2 Results

We found that the best LoVT-uni-gauss method outperforms LoVT /wo local on 10 of 18 tasks and CLIP on 9 tasks. On 6 tasks it could even outperform LoVT, showing that LoVT-uni-gauss is competitive with LoVT. If we choose the local temperature τ' individually per downstream task (out of two studied temperatures), the LoVT-uni-gauss method outperforms LoVT /wo local on 12, CLIP on 13 and LoVT on 7 of 18 tasks. This highlights the importance of the local temperature τ' and indicates that different tasks require different strengths of uniformity.

In order to study the effect of local uniformity in general, without the restriction to a single uniformity loss variant, we now consider both uniformity variants, uni-gauss and uni-xent. On 12 of 18 tasks (on 10 tasks by more then the 95%-confidence interval) uniformity based methods (uni-gauss or uni-xent) outperform all studied image-text methods without local losses (i.e. ConVIRT, CLIP, and LoVT /wo local). These results indicate that with well-tuned hyperparameters, uniformity-based methods outperform image-text methods that only rely on global losses on the large majority of studied tasks, even significantly on most of them. However, on 9 of 18 tasks LoVT outperforms all uniformity-based methods, from which we conclude that on some tasks the additional local alignment components of LoVT are still beneficial.

We also study how good performance of our uniformity-based methods correlate with good performance of LoVT when comparing them against methods without local losses. In Tab. 1 we therefore consider the cases (i.e. tasks) where uniformity-based methods (uni-gauss or uni-xent) are i) better than CLIP, ii) better than LoVT, iii) better than LoVT /wo local, and iv) better than all studied methods. For each of these cases we then compare on how many of them a) LoVT or CLIP is better and b) LoVT /wo local is better or worse than LoVT. We observe that on tasks where uniformity-based methods are better than CLIP or better than LoVT /wo local, LoVT often also outperforms CLIP and LoVT /wo local, indicating that local uniformity plays an important role in the success of LoVT. For detailed results we refer to Appendix E and for a comparison of the effective uniformity of representations we refer to Appendix B.

3.3 Conclusion

We studied the relationship of global and local contrastive losses of image-text methods for localized medical downstream tasks. Considering LoVT's loss functions, we found that the alignment components of the global and local losses are highly related and proposed to drop the local alignment components. We found that the distribution priors of the losses act complementary and proposed a local uniformity loss replacing the local distribution priors while keeping the global priors. Empirically, we found that our proposed method typically works well on chest X-ray tasks where local losses improve the results, indicating that local uniformity plays an important role when pretraining for localized tasks. On some tasks however the uniformity-based approaches are still outperformed by LoVT, suggesting that local contrastive losses cannot fully be replaced by local uniformity losses. We hope that our findings inspire future research to further improve pretraining on localized tasks.

References

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv: 2103.00020*, 2021.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D. Manning, and Curtis P. Langlotz. Contrastive learning of medical visual representations from paired images and text. *arXiv preprint arXiv: 2010.00747*, 2020.

Philip Müller, Georgios Kaissis, Congyu Zou, and Daniel Rueckert. Radiological reports improve pre-training for localized imaging tasks on chest x-rays. In *[to be published at] MICCAI*, 2022a.

Philip Müller, Georgios Kaissis, Congyu Zou, and Daniel Rueckert. Joint learning of localized representations from medical images and reports. In *[to be published at] ECCV, arXiv: 2112.02889*, 2022b.

Ruizhi Liao, Daniel Moyer, Miriam Cha, Keegan Quigley, Seth Berkowitz, Steven Horng, Polina Golland, and William M. Wells. Multimodal representation learning via maximization of local mutual information. In *MICCAI*, 2021.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *ICML*, 2020.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.

A. Johnson, T. Pollard, S. Berkowitz, et al. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Sci Data*, 6(317), 2019a. doi: https://doi.org/10.1038/s41597-019-0322-0.

A. Johnson, T. Pollard, R. Mark, S. Berkowitz, and S. Horng. Mimic-cxr database (version 2.0.0). PhysioNet, 2019b.

A. Johnson, M. Lungren, Y. Peng, et al. Mimic-cxr-jpg - chest radiographs with structured labels (version 2.0.0). PhysioNet, 2019c.

Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. *arXiv preprint arXiv: 2006.06666*, 2020.

Mert Sariyildiz, Julien Perez, and Diane Larlus. Learning visual representations with caption annotations. In *ECCV*, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. doi: 10.1109/CVPR.2016.90.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *AAAI*, pages 590–597, 2019.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In *NeurIPS*, 2020.

Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. *arXiv preprint arXiv: 2011.10043*, 2020.

A Relation to other Image-text Methods

While our study focuses on LoVT's loss functions, we now shortly explain the relation to other text-supervised methods.

Comparison to ConVIRT, CLIP, and ALIGN ConVIRT [Zhang et al., 2020], CLIP [Radford et al., 2021], and ALIGN [Jia et al., 2021] all follow a similar framework as LoVT but they do not align local representations explicitly, i.e. they only optimize $\mathcal{L}_{\text{global-align}}$ and $\mathcal{L}_{\text{global-dist}}$ but none of the local loss components. However, all of these methods consider only a single sentence per sample (in the case of reports, they randomly sample a single sentence from it). This means that $\mathcal{L}_{\text{global-align}}$ effectively aligns local report representations (i.e. sentences) with global image representations. However, $\mathcal{L}_{\text{global-dist}}$ still imposes a uniform distribution over the dataset and there is (unlike in LoVT) no component that enforces (local) representations to be uniformly distributed within each sample. Additionally, only CLIP uses (like LoVT) attention pooling to compute image representations while the other methods use average pooling.

Comparison to local-mi In the local-MI [Liao et al., 2021] work mutual information is maximized between (randomly sampled) sentence representations and the best matching image region representations. This should therefore have a similar effect as minimizing the local alignment components $\mathcal{L}^{\mathcal{I}}_{\text{local-align}}$ and $\mathcal{L}^{\mathcal{R}}_{\text{local-align}}$. However, when estimating the mutual information they consider the whole dataset. While it is not explicitly stated whether they also include other (non-matching) region-sentence pairs from the same sample during mutual information estimation, it is probable that their mutual information estimation imposes a distribution prior that is rather similar to $\mathcal{L}_{\text{global-dist}}$ than to the local distribution priors $\mathcal{L}^{\mathcal{I}}_{\text{local-dist}}$ and $\mathcal{L}^{\mathcal{R}}_{\text{local-dist}}$.

Comparison to VirTex and ICMLM VirTex [Desai and Johnson, 2020] and ICMLM [Sariyildiz et al., 2020] follow a different framework. Instead of using contrastive losses for pretraining, they use generative objectives by generating captions for images. While generative objectives may implicitly enforce alignment and distributions priors in order to achieve good generation performance, they are not enforced explicitly. Above all, image region representations are not enforced (at least not explicitly) to be distributed uniformly within each sample.

B Comparison of Uniformity

Fig. 1 shows the uniformity of local (i.e. scan region or report sentence) and global (i.e. scan or report) representations of different models. Local uniformity is measured using $-\mathcal{L}^{\mathcal{I}}_{\text{uni-gauss}}$ and $-\mathcal{L}^{\mathcal{R}}_{\text{uni-gauss}}$ on the local representations $\boldsymbol{y}^{\mathcal{I}}_{i,k}$ and $\boldsymbol{y}^{\mathcal{R}}_{i,m}$ (before projection), respectively, while the global uniformity is measured using the negative uniformity loss as defined in Wang and Isola [2020] on the scan $\bar{\boldsymbol{y}}^{\mathcal{I}}_i$ and report $\bar{\boldsymbol{y}}^{\mathcal{R}}_i$ representations (again before projection). Our local uniformity-based methods have, as expected, much larger per-sample uniformities (especially compared to LoVT /wo local), while at the same time having (slightly) smaller global uniformities. Larger temperatures increase this effect. Also, the effect is more present in the uni-xent methods compared to the uni-gauss methods.

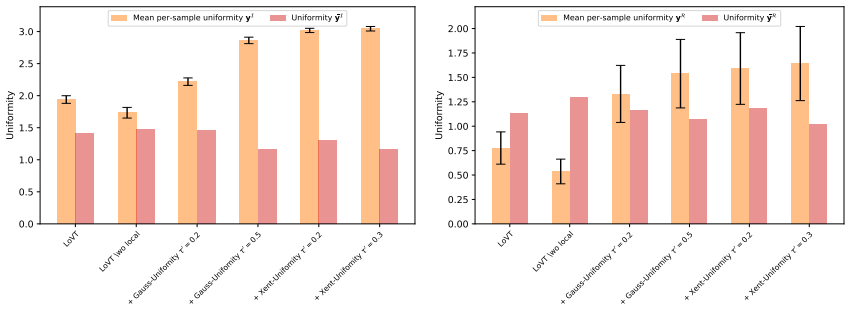

Figure 1: Uniformity of global and per-sample uniformity of local representations. **Left:** Image ($\bar{\boldsymbol{y}}_i^{\mathcal{I}}$) and image region ($\boldsymbol{y}_{i,k}^{\mathcal{I}}$) representations. **Right:** Report ($\bar{\boldsymbol{y}}_i^{\mathcal{R}}$) and report sentence ($\boldsymbol{y}_{i,m}^{\mathcal{R}}$) representations. The models were trained on 30% of frontal MIMIC-CXR and then evaluated on the whole test set.

C Detailed Derivations

C.1 Image-text Contrastive Framework of LoVT

Most image-text contrastive works [Radford et al., 2021, Jia et al., 2021, Zhang et al., 2020, Müller et al., 2022b, Liao et al., 2021] follow a similar framework. They use two independent encoders, one for encoding images and one for encoding the text (e.g. radiological reports). In the case of LoVT [Müller et al., 2022b], ResNet50 [He et al., 2016] is used to encode the images and BERT [Devlin et al., 2019] is used to encode the full reports. Each image $\boldsymbol{x}_i^{\mathcal{I}}$ is encoded into $K = H \times W$ (we use $K = 7 \times 7$) region representations $\boldsymbol{y}_{i,k}^{\mathcal{I}}$, taking the last feature map of the ResNet50. Similarly, each report $\boldsymbol{x}_i^{\mathcal{R}}$ is encoded into M_i sentence representations $\boldsymbol{y}_{i,m}^{\mathcal{R}}$ by max-pooling over all the token representations (encoded by BERT) of each sentence. The global (i.e. per-sample) representations of images $\bar{\boldsymbol{y}}_i^{\mathcal{I}}$ and reports $\bar{\boldsymbol{y}}_i^{\mathcal{R}}$ are each computed by an attention pooling layer on the region and sentence representations, respectively. All the local and global representations are then projected independently using (non-shared) projection heads based on multi-layer perceptrons (MLPs). We denote the local projected representations by $\boldsymbol{z}_{i,k}^{\mathcal{I}}$ and $\boldsymbol{z}_{i,m}^{\mathcal{R}}$, and the global representations by $\bar{\boldsymbol{z}}_i^{\mathcal{I}}$ and $\bar{\boldsymbol{z}}_i^{\mathcal{R}}$. LoVT additionally computes cross-modality representations $\boldsymbol{z}_{i,k}^{\mathcal{R}\to\mathcal{I}}$ and $\boldsymbol{z}_{i,m}^{\mathcal{I}\to\mathcal{R}}$ using an attention layer where the local representations of one modality attend to the representations of the other. These representations are then aligned using LoVT's loss function, which we will recapitulate in the following section.

C.2 Definition of LoVT's Loss Function

We will denote LoVT's loss function as $\mathcal{L}_{\text{LoVT}}$. It consists of three parts, the global contrasive loss $\mathcal{L}_{\text{global}}$, the local contrastive loss for image regions (aligned with cross-modality sentence representations) $\mathcal{L}_{\text{local-image}}$ and the local contrastive loss for report sentences (aligned with cross-modality region representations) $\mathcal{L}_{\text{local-report}}$:

$$\mathcal{L}_{\text{LoVT}} = \gamma \cdot \mathcal{L}_{\text{global}} + \mu \cdot \mathcal{L}_{\text{local-image}} + \nu \cdot \mathcal{L}_{\text{local-report}}\,, \tag{5}$$

where γ, μ, and ν are loss coefficients (hyperparameters).

The global contrastive loss $\mathcal{L}_{\text{global}}$ is applied to the global representations $\bar{\boldsymbol{z}}_i^{\mathcal{I}}$ and $\bar{\boldsymbol{z}}_i^{\mathcal{R}}$ and is defined as follows:

$$\ell_{\text{global}}^{\mathcal{I}\|\mathcal{R}} = -\log \frac{e^{\cos\left(\bar{\boldsymbol{z}}_i^{\mathcal{I}}, \bar{\boldsymbol{z}}_i^{\mathcal{R}}\right)/\tau}}{\sum_j e^{\cos\left(\bar{\boldsymbol{z}}_i^{\mathcal{I}}, \bar{\boldsymbol{z}}_j^{\mathcal{R}}\right)/\tau}} \tag{6}$$

$$\ell_{\text{global}}^{\mathcal{R}\|\mathcal{I}} = -\log \frac{e^{\cos\left(\bar{\boldsymbol{z}}_i^{\mathcal{R}}, \bar{\boldsymbol{z}}_i^{\mathcal{I}}\right)/\tau}}{\sum_j e^{\cos\left(\bar{\boldsymbol{z}}_i^{\mathcal{R}}, \bar{\boldsymbol{z}}_j^{\mathcal{I}}\right)/\tau}} \tag{7}$$

$$\mathcal{L}_{\text{global}} = \frac{1}{N}\sum_{i=1}^{N}\left[\lambda \cdot \ell_{\text{global}}^{\mathcal{I}\|\mathcal{R}} + (1-\lambda)\cdot \ell_{\text{global}}^{\mathcal{R}\|\mathcal{I}}\right], \tag{8}$$

where τ is the global temperatures, $\lambda \in [0,1]$ is a hyperparameter, and N denotes the batch size.

The local contrastive loss $\mathcal{L}_{\text{local-image}}$ for image regions is applied to $\boldsymbol{z}_{i,k}^{\mathcal{I}}$ and $\boldsymbol{z}_{i,k}^{\mathcal{R}\to\mathcal{I}}$ and is defined as:

$$\ell_{\text{local-image}}^{\mathcal{I}\|\mathcal{R}\to\mathcal{I}} = -\sum_{l=1}^{K} p_{k,l}^{\mathcal{I}} \log \frac{e^{\cos\left(\boldsymbol{z}_{i,k}^{\mathcal{I}}, \boldsymbol{z}_{i,l}^{\mathcal{R}\to\mathcal{I}}\right)/\tau'}}{\sum_{k'} e^{\cos\left(\boldsymbol{z}_{i,k}^{\mathcal{I}}, \boldsymbol{z}_{i,k'}^{\mathcal{R}\to\mathcal{I}}\right)/\tau'}} \tag{9}$$

$$\ell_{\text{local-image}}^{\mathcal{R}\to\mathcal{I}\|\mathcal{I}} = -\sum_{l=1}^{K} p_{k,l}^{\mathcal{I}} \log \frac{e^{\cos\left(\boldsymbol{z}_{i,k}^{\mathcal{R}\to\mathcal{I}}, \boldsymbol{z}_{i,l}^{\mathcal{I}}\right)/\tau'}}{\sum_{k'} e^{\cos\left(\boldsymbol{z}_{i,k}^{\mathcal{R}\to\mathcal{I}}, \boldsymbol{z}_{i,k'}^{\mathcal{I}}\right)/\tau'}} \tag{10}$$

$$\mathcal{L}_{\text{local-image}} = \frac{1}{2N}\sum_{i=1}^{N}\sum_{k=1}^{K} w_{i,k}^{\mathcal{I}} \cdot \left[\ell_{\text{local-image}}^{\mathcal{I}\|\mathcal{R}\to\mathcal{I}} + \ell_{\text{local-image}}^{\mathcal{R}\to\mathcal{I}\|\mathcal{I}}\right], \tag{11}$$

where τ' is the local temperature, $w_{i,k}^{\mathcal{I}}$ is the weight for image region k (computed based on the image of sample i) and $p_{k,l}^{\mathcal{I}}$ is the positiveness weight for the region pair (k,l) (computed based on their spatial distance).

The local contrastive loss $\mathcal{L}_{\text{local-report}}$ for report sentences is applied to $\boldsymbol{z}_{i,m}^{\mathcal{R}}$ and $\boldsymbol{z}_{i,m}^{\mathcal{I}\to\mathcal{R}}$ and is defined as:

$$\ell_{\text{local-report}}^{\mathcal{R}\|\mathcal{I}\to\mathcal{R}} = -\log \frac{e^{\cos\left(\boldsymbol{z}_{i,m}^{\mathcal{R}}, \boldsymbol{z}_{i,m}^{\mathcal{I}\to\mathcal{R}}\right)/\tau'}}{\sum_{m'} e^{\cos\left(\boldsymbol{z}_{i,m}^{\mathcal{R}}, \boldsymbol{z}_{i,m'}^{\mathcal{I}\to\mathcal{R}}\right)/\tau'}} \tag{12}$$

$$\ell_{\text{local-report}}^{\mathcal{I}\to\mathcal{R}\|\mathcal{R}} = -\log \frac{e^{\cos\left(\boldsymbol{z}_{i,m}^{\mathcal{I}\to\mathcal{R}}, \boldsymbol{z}_{i,m}^{\mathcal{R}}\right)/\tau'}}{\sum_{m'} e^{\cos\left(\boldsymbol{z}_{i,m}^{\mathcal{I}\to\mathcal{R}}, \boldsymbol{z}_{i,m'}^{\mathcal{R}}\right)/\tau'}} \tag{13}$$

$$\mathcal{L}_{\text{local-report}} = \frac{1}{2N}\sum_{i=1}^{N}\sum_{m=1}^{M_i} w_{i,m}^{\mathcal{R}} \cdot \left[\ell_{\text{local-report}}^{\mathcal{R}\|\mathcal{I}\to\mathcal{R}} + \ell_{\text{local-report}}^{\mathcal{I}\to\mathcal{R}\|\mathcal{R}}\right], \tag{14}$$

where $w_{i,m}^{\mathcal{R}}$ is the weight for sentence m (computed based on the report of sample i).

For a detailed explanation of the components and their intuition we refer to the original publication of LoVT [Müller et al., 2022b].

C.3 Decomposition of the Contrastive Losses of LoVT

We now start our analysis with the decomposition of LoVT's loss function. Each of the three parts of LoVT's loss function (i.e. $\mathcal{L}_{\text{global}}$, $\mathcal{L}_{\text{local-image}}$, and $\mathcal{L}_{\text{local-report}}$) are NTXent-based [Chen et al., 2020] contrastive loss functions and following previous works [Wang and Isola, 2020] can therefore each be decomposed into an *alignment* and a *distribution prior* component. We decompose $\mathcal{L}_{\text{global}}$:

$$\mathcal{L}_{\text{global}} = \mathcal{L}_{\text{global-align}} + \mathcal{L}_{\text{global-dist}} \tag{15}$$

into the global alignment component

$$\mathcal{L}_{\text{global-align}} = -\frac{1}{\tau}\frac{1}{N}\sum_{i=1}^{N}\cos\left(\bar{\boldsymbol{z}}_i^{\mathcal{I}}, \bar{\boldsymbol{z}}_i^{\mathcal{R}}\right) \tag{16}$$

and the global distribution prior

$$\begin{aligned}\mathcal{L}_{\text{global-dist}} = \lambda\frac{1}{N}\sum_{i=1}^{N}\log\sum_{j=1}^{N}\exp\left(\frac{\cos\left(\bar{\boldsymbol{z}}_i^{\mathcal{I}}, \bar{\boldsymbol{z}}_j^{\mathcal{R}}\right)}{\tau}\right) \\ + (1-\lambda)\frac{1}{N}\sum_{i=1}^{N}\log\sum_{j=1}^{N}\exp\left(\frac{\cos\left(\bar{\boldsymbol{z}}_i^{\mathcal{R}}, \bar{\boldsymbol{z}}_j^{\mathcal{I}}\right)}{\tau}\right)\end{aligned} \tag{17}$$

Similarly, we can decompose the local losses $\mathcal{L}_{\text{local-image}}$ and $\mathcal{L}_{\text{local-report}}$:

$$\mathcal{L}_{\text{local-image}} = \mathcal{L}^{\mathcal{I}}_{\text{local-align}} + \mathcal{L}^{\mathcal{I}}_{\text{local-dist}} \tag{18}$$

$$\cdot\mathcal{L}_{\text{local-report}} = \mathcal{L}^{\mathcal{R}}_{\text{local-align}} + \mathcal{L}^{\mathcal{R}}_{\text{local-dist}} \tag{19}$$

into the local alignment component for scan regions

$$\mathcal{L}^{\mathcal{I}}_{\text{local-align}} = -\frac{1}{\tau'}\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{l=1}^{K}\frac{w^{\mathcal{I}}_{i,k}p^{\mathcal{I}}_{k,l} + w^{\mathcal{I}}_{i,l}p^{\mathcal{I}}_{l,k}}{2}\cos\left(\boldsymbol{z}^{\mathcal{I}}_{i,k}, \boldsymbol{z}^{\mathcal{R}\to\mathcal{I}}_{i,l}\right) \tag{20}$$

and the alignment for report sentences

$$\mathcal{L}^{\mathcal{R}}_{\text{local-align}} = -\frac{1}{\tau'}\frac{1}{N}\sum_{i=1}^{N}\sum_{m=1}^{M_i} w^{\mathcal{R}}_{i,m}\cos\left(\boldsymbol{z}^{\mathcal{R}}_{i,m}, \boldsymbol{z}^{\mathcal{I}\to\mathcal{R}}_{i,m}\right) \tag{21}$$

as well as into the distribution prior for scan regions

$$\begin{aligned}\mathcal{L}^{\mathcal{I}}_{\text{local-dist}} = &\frac{1}{2N}\sum_{i=1}^{N}\sum_{k=1}^{K} w^{\mathcal{I}}_{i,k}\log\sum_{k'=1}^{K}\exp\left(\frac{\cos\left(\boldsymbol{z}^{\mathcal{I}}_{i,k}, \boldsymbol{z}^{\mathcal{R}\to\mathcal{I}}_{i,k'}\right)}{\tau'}\right)\\ &+\frac{1}{2N}\sum_{i=1}^{N}\sum_{k=1}^{K} w^{\mathcal{I}}_{i,k}\log\sum_{k'=1}^{K}\exp\left(\frac{\cos\left(\boldsymbol{z}^{\mathcal{R}\to\mathcal{I}}_{i,k}, \boldsymbol{z}^{\mathcal{I}}_{i,k'}\right)}{\tau'}\right)\end{aligned} \tag{22}$$

and the prior for report sentences

$$\begin{aligned}\mathcal{L}^{\mathcal{R}}_{\text{local-dist}} = &\frac{1}{2N}\sum_{i=1}^{N}\sum_{m=1}^{M_i} w^{\mathcal{R}}_{i,m}\log\sum_{m'=1}^{M_i}\exp\left(\frac{\cos\left(\boldsymbol{z}^{\mathcal{R}}_{i,m}, \boldsymbol{z}^{\mathcal{I}\to\mathcal{R}}_{i,m'}\right)}{\tau'}\right)\\ &+\frac{1}{2N}\sum_{i=1}^{N}\sum_{m=1}^{M_i} w^{\mathcal{R}}_{i,m}\log\sum_{m'=1}^{M_i}\exp\left(\frac{\cos\left(\boldsymbol{z}^{\mathcal{I}\to\mathcal{R}}_{i,m}, \boldsymbol{z}^{\mathcal{R}}_{i,m'}\right)}{\tau'}\right).\end{aligned} \tag{23}$$

In the following sections we will now analyze each of these components individually.

C.4 Alignment Components

We now study the three alignment components $\mathcal{L}_{\text{global-align}}$, $\mathcal{L}^{\mathcal{I}}_{\text{local-align}}$, and $\mathcal{L}^{\mathcal{R}}_{\text{local-align}}$. The global alignment component $\mathcal{L}_{\text{global-align}}$ forces (global) image and report representations from the same sample to be aligned by maximizing their cosine similarity. The local alignment components $\mathcal{L}^{\mathcal{I}}_{\text{local-align}}$ and $\mathcal{L}^{\mathcal{R}}_{\text{local-align}}$ on the other hand align local representations of image regions or report sentences with local cross-modality representations computed using attention on local representations from the other modality. The individual (local) cosine similarities are aggregated as weighted sums per modality and then maximized.

We will now study how the global alignment component is related to the local alignment components and will therefore make some simplifying assumptions. First, we simplify the attention pooling operation used to compute global representations and replace it by a weighted sum (ignoring that attention pooling uses multiple heads and linear projections). Second, we ignore that local representations $\boldsymbol{z}^{\mathcal{I}}_{i,k}$ and $\boldsymbol{z}^{\mathcal{R}}_{i,m}$ are projected independently from the global representations $\bar{\boldsymbol{z}}^{\mathcal{I}}_i$ and $\bar{\boldsymbol{z}}^{\mathcal{R}}_i$. Instead, we assume that $\bar{\boldsymbol{z}}^{\mathcal{I}}_i$ and $\bar{\boldsymbol{z}}^{\mathcal{R}}_i$ are computed as a weighted sum of local representations:

$$\bar{\boldsymbol{z}}^{\mathcal{I}}_i = \sum_{k=1}^{K} w^{\mathcal{I}}_{i,k}\boldsymbol{z}^{\mathcal{I}}_{i,k} \qquad \bar{\boldsymbol{z}}^{\mathcal{R}}_i = \sum_{m=1}^{M_i} w^{\mathcal{R}}_{i,m}\boldsymbol{z}^{\mathcal{R}}_{i,m} \tag{24}$$

Similarly, we simplify the computation of the cross-modality representations $\boldsymbol{z}^{\mathcal{I}\to\mathcal{R}}_{i,m}$ and $\boldsymbol{z}^{\mathcal{R}\to\mathcal{I}}_{i,k}$ by ignoring the linear projections of the attention mechanism used in LoVT and assume:

$$\boldsymbol{z}^{\mathcal{I}\to\mathcal{R}}_{i,m} = \sum_{k=1}^{K}\alpha^{\mathcal{I}\to\mathcal{R}}_{i,m,k}\boldsymbol{z}^{\mathcal{I}}_{i,k} \qquad \boldsymbol{z}^{\mathcal{R}\to\mathcal{I}}_{i,k} = \sum_{m=1}^{M_i}\alpha^{\mathcal{R}\to\mathcal{I}}_{i,k,m}\boldsymbol{z}^{\mathcal{R}}_{i,m} \tag{25}$$

We also ignore the normalization terms of the cosine similarity, i.e. we treat all cosine similarities as dot products. Note that this is equal to assuming that $\boldsymbol{z}^{\mathcal{I}}_{i,k}$, $\boldsymbol{z}^{\mathcal{R}}_{i,m}$, $\bar{\boldsymbol{z}}^{\mathcal{I}}_{i}$, $\bar{\boldsymbol{z}}^{\mathcal{R}}_{i}$, $\boldsymbol{z}^{\mathcal{I}\to\mathcal{R}}_{i,m}$ and $\boldsymbol{z}^{\mathcal{R}\to\mathcal{I}}_{i,k}$ are unit vectors which however does not hold considering Eqs. (24) and (25).

Using these simplifications we can rewrite the alignment losses. The global alignment loss is rewritten as

$$\mathcal{L}_{\text{global-align}} = -\frac{1}{\tau}\frac{1}{N}\sum_{i=1}^{N}\text{dot}\left(\sum_{k=1}^{K} w^{\mathcal{I}}_{i,k}\boldsymbol{z}^{\mathcal{I}}_{i,k}, \sum_{m=1}^{M_i} w^{\mathcal{R}}_{i,m}\boldsymbol{z}^{\mathcal{R}}_{i,m}\right) \tag{26}$$

$$= -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{m=1}^{M_i}\frac{w^{\mathcal{I}}_{i,k}w^{\mathcal{R}}_{i,m}}{\tau}\text{dot}\left(\boldsymbol{z}^{\mathcal{I}}_{i,k}, \boldsymbol{z}^{\mathcal{R}}_{i,m}\right) \tag{27}$$

$$= -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{m=1}^{M_i}\xi^{\mathcal{G}}_{i,k,m}\text{dot}\left(\boldsymbol{z}^{\mathcal{I}}_{i,k}, \boldsymbol{z}^{\mathcal{R}}_{i,m}\right)\ , \tag{28}$$

the local scan alignment as

$$\mathcal{L}^{\mathcal{I}}_{\text{local-align}} = -\frac{1}{\tau'}\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{l=1}^{K}\frac{w^{\mathcal{I}}_{i,k}p^{\mathcal{I}}_{k,l} + w^{\mathcal{I}}_{i,l}p^{\mathcal{I}}_{l,k}}{2}\text{dot}\left(\boldsymbol{z}^{\mathcal{I}}_{i,k}, \sum_{m=1}^{M_i}\alpha^{\mathcal{R}\to\mathcal{I}}_{i,l,m}\boldsymbol{z}^{\mathcal{R}}_{i,m}\right) \tag{29}$$

$$= -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{m=1}^{M_i}\frac{\sum_{l=1}^{K}(w^{\mathcal{I}}_{i,k}p^{\mathcal{I}}_{k,l} + w^{\mathcal{I}}_{i,l}p^{\mathcal{I}}_{l,k})\cdot\alpha^{\mathcal{R}\to\mathcal{I}}_{i,l,m}}{\tau'}\text{dot}\left(\boldsymbol{z}^{\mathcal{I}}_{i,k}, \boldsymbol{z}^{\mathcal{R}}_{i,m}\right) \tag{30}$$

$$= -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{m=1}^{M_i}\xi^{\mathcal{I}}_{i,k,m}\text{dot}\left(\boldsymbol{z}^{\mathcal{I}}_{i,k}, \boldsymbol{z}^{\mathcal{R}}_{i,m}\right)\ , \tag{31}$$

and the local report alignment is rewritten as

$$\mathcal{L}^{\mathcal{R}}_{\text{local-align}} = -\frac{1}{\tau'}\frac{1}{N}\sum_{i=1}^{N}\sum_{m=1}^{M_i}w^{\mathcal{R}}_{i,m}\text{dot}\left(\boldsymbol{z}^{\mathcal{R}}_{i,m}, \sum_{m=1}^{M_i}\alpha^{\mathcal{I}\to\mathcal{R}}_{i,m,k}\boldsymbol{z}^{\mathcal{I}}_{i,k}\right) \tag{32}$$

$$= -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{m=1}^{M_i}\frac{w^{\mathcal{R}}_{i,m}\alpha^{\mathcal{I}\to\mathcal{R}}_{i,m,k}}{\tau'}\text{dot}\left(\boldsymbol{z}^{\mathcal{R}}_{i,m}, \boldsymbol{z}^{\mathcal{I}}_{i,k}\right) \tag{33}$$

$$= -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{m=1}^{M_i}\xi^{\mathcal{R}}_{i,k,m}\text{dot}\left(\boldsymbol{z}^{\mathcal{R}}_{i,m}, \boldsymbol{z}^{\mathcal{I}}_{i,k}\right)\ . \tag{34}$$

Comparing Eqs. (28), (31) and (34), we realize that they are all special cases of the same general form of alignment loss that maximizes a weighted sum of all possible pairs of cosine similarities between two local representations from different modalities. The only difference between the three losses is how the weights of the sum, i.e. $\xi_{i,k,m}$, are defined.

We also realize that $\xi^{\mathcal{I}}_{i,k,m}$ and $\xi^{\mathcal{R}}_{i,k,m}$ are, unlike $\xi^{\mathcal{G}}_{i,k,m}$, not separable into components containing only information from a single modality (image or report). Thus, $\mathcal{L}^{\mathcal{I}}_{\text{local-align}}$ and $\mathcal{L}^{\mathcal{R}}_{\text{local-align}}$ cannot be rewritten in the original form of $\mathcal{L}_{\text{global-align}}$. If we assume that $\alpha^{\mathcal{R}\to\mathcal{I}}_{i,l,m}$ and $\alpha^{\mathcal{I}\to\mathcal{R}}_{i,m,k}$ are not computed as in LoVT but are instead separable by modality, i.e. $\alpha^{\mathcal{R}\to\mathcal{I}}_{i,l,m} = \tilde{\alpha}^{\mathcal{R}\to\mathcal{I}}_{i,l}\hat{\alpha}^{\mathcal{R}\to\mathcal{I}}_{i,m}$ and $\alpha^{\mathcal{I}\to\mathcal{R}}_{i,m,k} = \tilde{\alpha}^{\mathcal{I}\to\mathcal{R}}_{i,m}\hat{\alpha}^{\mathcal{I}\to\mathcal{R}}_{i,k}$, then we can rewrite $\xi^{\mathcal{I}}_{i,k,m}$ and $\xi^{\mathcal{R}}_{i,k,m}$ as

$$\xi^{\mathcal{I}}_{i,k,m} = \frac{1}{\tau'}\sum_{l=1}^{K}(w^{\mathcal{I}}_{i,k}p^{\mathcal{I}}_{k,l} + w^{\mathcal{I}}_{i,l}p^{\mathcal{I}}_{l,k})\cdot\tilde{\alpha}^{\mathcal{R}\to\mathcal{I}}_{i,l}\hat{\alpha}^{\mathcal{R}\to\mathcal{I}}_{i,m} \tag{35}$$

$$\xi^{\mathcal{R}}_{i,k,m} = \frac{1}{\tau'}w^{\mathcal{R}}_{i,m}\tilde{\alpha}^{\mathcal{I}\to\mathcal{R}}_{i,m}\hat{\alpha}^{\mathcal{I}\to\mathcal{R}}_{i,k} \tag{36}$$

and therefore as

$$\xi^{\mathcal{I}}_{i,k,m} = \tilde{\xi}^{\mathcal{I}}_{i,k}\hat{\xi}^{\mathcal{I}}_{i,m} \qquad \xi^{\mathcal{R}}_{i,k,m} = \tilde{\xi}^{\mathcal{R}}_{i,k}\hat{\xi}^{\mathcal{R}}_{i,m} \tag{37}$$

with

$$\tilde{\xi}^{\mathcal{I}}_{i,k} = \sqrt{\frac{1}{\tau'}} \sum_{l=1}^{K} (w^{\mathcal{I}}_{i,k} p^{\mathcal{I}}_{k,l} + w^{\mathcal{I}}_{i,l} p^{\mathcal{I}}_{l,k}) \cdot \tilde{\alpha}^{\mathcal{R}\to\mathcal{I}}_{i,l} \qquad \hat{\xi}^{\mathcal{I}}_{i,m} = \sqrt{\frac{1}{\tau'}} \hat{\alpha}^{\mathcal{R}\to\mathcal{I}}_{i,m} \tag{38}$$

$$\tilde{\xi}^{\mathcal{R}}_{i,k} = \sqrt{\frac{1}{\tau'}} \hat{\alpha}^{\mathcal{I}\to\mathcal{R}}_{i,k} \qquad \hat{\xi}^{\mathcal{R}}_{i,m} = \sqrt{\frac{1}{\tau'}} w^{\mathcal{R}}_{i,m} \tilde{\alpha}^{\mathcal{I}\to\mathcal{R}}_{i,m} \tag{39}$$

Using these definitions, we can write

$$\begin{aligned} \mathcal{L}^{\mathcal{I}}_{\text{local-align}} &= -\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} \sum_{m=1}^{M_i} \tilde{\xi}^{\mathcal{I}}_{i,k} \hat{\xi}^{\mathcal{I}}_{i,m} \text{dot}\left(\boldsymbol{z}^{\mathcal{I}}_{i,k}, \boldsymbol{z}^{\mathcal{R}}_{i,m}\right) \\ &= -\frac{1}{N} \sum_{i=1}^{N} \text{dot}\left(\sum_{k=1}^{K} \tilde{\xi}^{\mathcal{I}}_{i,k} \boldsymbol{z}^{\mathcal{I}}_{i,k}, \sum_{m=1}^{M_i} \hat{\xi}^{\mathcal{I}}_{i,m} \boldsymbol{z}^{\mathcal{R}}_{i,m} \right) \end{aligned} \tag{40}$$

and

$$\begin{aligned} \mathcal{L}^{\mathcal{R}}_{\text{local-align}} &= -\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} \sum_{m=1}^{M_i} \tilde{\xi}^{\mathcal{R}}_{i,k} \hat{\xi}^{\mathcal{R}}_{i,m} \text{dot}\left(\boldsymbol{z}^{\mathcal{I}}_{i,k}, \boldsymbol{z}^{\mathcal{R}}_{i,m}\right) \\ &= -\frac{1}{N} \sum_{i=1}^{N} \text{dot}\left(\sum_{k=1}^{K} \tilde{\xi}^{\mathcal{R}}_{i,k} \boldsymbol{z}^{\mathcal{I}}_{i,k}, \sum_{m=1}^{M_i} \hat{\xi}^{\mathcal{R}}_{i,m} \boldsymbol{z}^{\mathcal{R}}_{i,m} \right) \end{aligned} \tag{41}$$

Comparing Eqs. (40) and (41) with the global alignment loss as defined in Eq. (26), we see that they only differ in the computations of the weights in the weighted sums within the cosine similarity. We therefore conclude, that the non-separability of $\xi^{\mathcal{I}}_{i,k,m}$ and $\xi^{\mathcal{R}}_{i,k,m}$ is an important aspect when comparing the global with the local contrastive losses.

C.5 Distribution Prior Components

C.5.1 Comparison of the Distributions Priors

We now consider the distribution priors of the loss function decomposition. The global distribution component $\mathcal{L}_{\text{global-dist}}$ imposes a prior on the distribution of image and sentence representations such that they do not collapse to a constant value. Minimizing $\mathcal{L}_{\text{global-dist}}$ introduces contrast by pushing representations from both modalities away from each other, i.e. maximizing the distance between $\bar{\boldsymbol{z}}^{\mathcal{I}}_i$ and $\bar{\boldsymbol{z}}^{\mathcal{R}}_j$ for all i, j. Considering that $\mathcal{L}_{\text{global-align}}$ pushes the scan and report representations from the same sample i closer to each other, it can be expected that $\mathcal{L}_{\text{global-dist}}$ pushes representations from different samples ($i \neq j$) further apart than the representations from the same sample ($i = j$). It thus pushes representations (even from the same modality) further apart and therefore imposes a uniform distribution on the representations from each modality, i.e. it enforces both, $\bar{\boldsymbol{z}}^{\mathcal{I}}_i$ and $\bar{\boldsymbol{z}}^{\mathcal{R}}_i$, to be roughly uniformly distributed. To understand this, consider the extreme case where $\mathcal{L}_{\text{global-align}}$ is constant, i.e. $\cos(\bar{\boldsymbol{z}}^{\mathcal{I}}_i, \bar{\boldsymbol{z}}^{\mathcal{R}}_i) = c_i$ for each i. The only way to minimize $\mathcal{L}_{\text{global-dist}}$ is then to also push representations from the same modality (e.g. $\bar{\boldsymbol{z}}^{\mathcal{I}}_j$ for different samples $j \neq i$) further apart, i.e. to maximize $\cos(\bar{\boldsymbol{z}}^{\mathcal{I}}_i, \bar{\boldsymbol{z}}^{\mathcal{I}}_j)$ for $i \neq j$.

For comparing the global distribution prior $\mathcal{L}_{\text{global-dist}}$ to the local distribution priors $\mathcal{L}^{\mathcal{I}}_{\text{local-dist}}$ and $\mathcal{L}^{\mathcal{R}}_{\text{local-dist}}$, we first rewrite them by again assuming Eqs. (24) and (25) hold, such that we have

$$\begin{aligned} \mathcal{L}_{\text{global-dist}} = \lambda \frac{1}{N} \sum_{i=1}^{N} \log \sum_{j=1}^{N} \exp\left(\frac{\cos\left(\sum_{k=1}^{K} w^{\mathcal{I}}_{i,k} \boldsymbol{z}^{\mathcal{I}}_{i,k}, \sum_{m=1}^{M_i} w^{\mathcal{R}}_{j,m} \boldsymbol{z}^{\mathcal{R}}_{j,m} \right)}{\tau} \right) \\ + (1-\lambda) \frac{1}{N} \sum_{i=1}^{N} \log \sum_{j=1}^{N} \exp\left(\frac{\cos\left(\sum_{m=1}^{M_i} w^{\mathcal{R}}_{i,m} \boldsymbol{z}^{\mathcal{R}}_{i,m}, \sum_{k=1}^{K} w^{\mathcal{I}}_{j,s} \boldsymbol{z}^{\mathcal{I}}_{j,k} \right)}{\tau} \right). \end{aligned} \tag{42}$$

Similarly, we rewrite $\mathcal{L}^{\mathcal{I}}_{\text{local-dist}}$ as

$$\begin{aligned}\mathcal{L}^{\mathcal{I}}_{\text{local-dist}} = &\frac{1}{2N}\sum_{i=1}^{N}\sum_{k=1}^{K} w^{\mathcal{I}}_{i,k} \log \sum_{k'=1}^{K} \exp\left(\frac{\cos\left(\boldsymbol{z}^{\mathcal{I}}_{i,k}, \sum_{m=1}^{M_i} \alpha^{\mathcal{R}\to\mathcal{I}}_{i,k',m} \boldsymbol{z}^{\mathcal{R}}_{i,m}\right)}{\tau'}\right) \\ &+ \frac{1}{2N}\sum_{i=1}^{N}\sum_{k=1}^{K} w^{\mathcal{I}}_{i,k} \log \sum_{k'=1}^{K} \exp\left(\frac{\cos\left(\sum_{m=1}^{M_i} \alpha^{\mathcal{R}\to\mathcal{I}}_{i,k,m} \boldsymbol{z}^{\mathcal{R}}_{i,m}, \boldsymbol{z}^{\mathcal{I}}_{i,k'}\right)}{\tau'}\right)\end{aligned} \tag{43}$$

and $\mathcal{L}^{\mathcal{R}}_{\text{local-dist}}$ as

$$\begin{aligned}\mathcal{L}^{\mathcal{R}}_{\text{local-dist}} = &\frac{1}{2N}\sum_{i=1}^{N}\sum_{m=1}^{M_i} w^{\mathcal{R}}_{i,m} \log \sum_{m'=1}^{M_i} \exp\left(\frac{\cos\left(\boldsymbol{z}^{\mathcal{R}}_{i,m}, \sum_{k=1}^{K} \alpha^{\mathcal{I}\to\mathcal{R}}_{i,m',k} \boldsymbol{z}^{\mathcal{I}}_{i,k}\right)}{\tau'}\right) \\ &+ \frac{1}{2N}\sum_{i=1}^{N}\sum_{m=1}^{M_i} w^{\mathcal{R}}_{i,m} \log \sum_{m'=1}^{M_i} \exp\left(\frac{\cos\left(\sum_{k=1}^{K} \alpha^{\mathcal{I}\to\mathcal{R}}_{i,m,k} \boldsymbol{z}^{\mathcal{I}}_{i,k}, \boldsymbol{z}^{\mathcal{R}}_{i,m'}\right)}{\tau'}\right).\end{aligned} \tag{44}$$

Comparing Eq. (42) with Eq. (43) and Eq. (44), we realize that they follow a similar form as they both minimize the (weighted) average of `logsumexp`-aggregations over pairwise cosine-similarities between local representations $\boldsymbol{z}^{\mathcal{I}}_{i,k}$ and $\boldsymbol{z}^{\mathcal{R}}_{i,m}$ (or weighted sums of them). However while in $\mathcal{L}_{\text{global-dist}}$ the `logsumexp`-aggregation is done over the samples in the batch (i.e. we sum over N samples), in $\mathcal{L}^{\mathcal{I}}_{\text{local-dist}}$ and $\mathcal{L}^{\mathcal{R}}_{\text{local-dist}}$ `logsumexp`-aggregation is done over the local representations of each sample (i.e. we sum over K or M_i regions or sentences, respectively, per sample). This means that while the $\mathcal{L}_{\text{global-dist}}$ loss imposes representations to be roughly uniformly distributed over the whole dataset, $\mathcal{L}^{\mathcal{I}}_{\text{local-dist}}$ and $\mathcal{L}^{\mathcal{R}}_{\text{local-dist}}$ impose uniform distributions of (local) representations over each sample, i.e. it pushes (local) representations within each sample away from each other and not across samples.

C.5.2 Cross-Entropy Uniformity Applied to each Modality Independently

We now propose to replace the local distribution component $\mathcal{L}^{\mathcal{I}}_{\text{local-dist}}$ and $\mathcal{L}^{\mathcal{R}}_{\text{local-dist}}$ by distribution components that impose a per-sample uniform distribution on each modality independently. Therefore, instead of using cosine similarities between (local) representations from different modalities, we apply the cosine similarities to pairs of (local) representations from the same modality. Additionally, we ignore the local weights $w^{\mathcal{I}}_{i,k}$ and $w^{\mathcal{R}}_{i,m}$ and use unweighted averages instead. We call the resulting local distributions components *cross-entropy uniformity (uni-xent)* as it is similar to the distribution component of the cross-entropy loss. We define the (local) cross-entropy uniformity for scans as:

$$\mathcal{L}^{\mathcal{I}}_{\text{uni-xent}} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{K}\sum_{k=1}^{K}\log\sum_{k'=1}^{K}\exp\left(-\frac{\cos(\boldsymbol{z}^{\mathcal{I}}_{i,k}, \boldsymbol{z}^{\mathcal{I}}_{i,k'})}{\tau'}\right) \tag{45}$$

and for reports as:

$$\mathcal{L}^{\mathcal{R}}_{\text{uni-xent}} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{M_i}\sum_{m=1}^{M_i}\log\sum_{m'=1}^{M_i}\exp\left(-\frac{\cos(\boldsymbol{z}^{\mathcal{R}}_{i,m}, \boldsymbol{z}^{\mathcal{R}}_{i,m'})}{\tau'}\right). \tag{46}$$

Like the local distribution components $\mathcal{L}^{\mathcal{I}}_{\text{local-dist}}$ and $\mathcal{L}^{\mathcal{R}}_{\text{local-dist}}$, the uniformity losses $\mathcal{L}^{\mathcal{I}}_{\text{uni-xent}}$ and $\mathcal{L}^{\mathcal{R}}_{\text{uni-xent}}$ minimize the cosine similarity of region (or sentence) representations within each sample and therefore push local representations within each sample away from each other. However, unlike $\mathcal{L}^{\mathcal{I}}_{\text{local-dist}}$ and $\mathcal{L}^{\mathcal{R}}_{\text{local-dist}}$, $\mathcal{L}^{\mathcal{I}}_{\text{uni-xent}}$ and $\mathcal{L}^{\mathcal{R}}_{\text{uni-xent}}$ do not have a repulsive effect between the modalities, i.e. they do not push regions and sentence representations away from each other. The global distribution prior $\mathcal{L}_{\text{global-dist}}$ has a similar repulsive effect and we therefore argue that we can replace $\mathcal{L}^{\mathcal{I}}_{\text{local-dist}}$ by $\mathcal{L}^{\mathcal{I}}_{\text{uni-xent}}$ and $\mathcal{L}^{\mathcal{R}}_{\text{local-dist}}$ by $\mathcal{L}^{\mathcal{R}}_{\text{uni-xent}}$ while still imposing similar prior distributions.

C.5.3 Pairwise Gaussian Potential

Following Wang and Isola [2020] we now study replacing the cross-entropy-based (local) uniformity losses $\mathcal{L}^{\mathcal{I}}_{\text{uni-xent}}$ and $\mathcal{L}^{\mathcal{R}}_{\text{uni-xent}}$ by losses based on pairwise Gaussian potentials. We therefore adapt the

uniformity loss proposed by Wang and Isola [2020] to impose uniform distribution within each sample instead of over the whole dataset. This leads to the *Gaussian uniformity (uni-gauss)* loss. For scans we denote it as $\mathcal{L}^{\mathcal{I}}_{\text{uni-gauss}}$ and define it as follows:

$$\mathcal{L}^{\mathcal{I}}_{\text{uni-gauss}} = \frac{1}{N}\sum_{i=1}^{N}\log\frac{1}{K^2}\sum_{k=1}^{K}\sum_{k'=1}^{K}\exp\left(\frac{\|\tilde{\boldsymbol{z}}^{\mathcal{I}}_{i,k} - \tilde{\boldsymbol{z}}^{\mathcal{I}}_{i,k'}\|_2^2}{\tau'}\right) \tag{47}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\log\frac{1}{K^2}\sum_{k=1}^{K}\sum_{k'=1}^{K}\exp\left(\frac{2 - 2\cdot\cos(\boldsymbol{z}^{\mathcal{I}}_{i,k}, \boldsymbol{z}^{\mathcal{I}}_{i,k'})}{\tau'}\right) \tag{48}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\log\frac{1}{K^2}\sum_{k=1}^{K}\sum_{k'=1}^{K}\exp\left(-\frac{\cos(\boldsymbol{z}^{\mathcal{I}}_{i,k}, \boldsymbol{z}^{\mathcal{I}}_{i,k'})}{\tau'}\right) + C\,, \tag{49}$$

where C is come constant and can therefore be dropped when minimizing $\mathcal{L}^{\mathcal{I}}_{\text{uni-gauss}}$, leading to our final definition:

$$\mathcal{L}^{\mathcal{I}}_{\text{uni-gauss}} = \frac{1}{N}\sum_{i=1}^{N}\log\frac{1}{K^2}\sum_{k=1}^{K}\sum_{k'=1}^{K}\exp\left(-\frac{\cos(\boldsymbol{z}^{\mathcal{I}}_{i,k}, \boldsymbol{z}^{\mathcal{I}}_{i,k'})}{\tau'}\right)\,. \tag{50}$$

Similarly, we define the uniformity loss for reports as

$$\mathcal{L}^{\mathcal{R}}_{\text{uni-gauss}} = \frac{1}{N}\sum_{i=1}^{N}\log\frac{1}{M_i^2}\sum_{m=1}^{M_i}\sum_{m'=1}^{M_i}\exp\left(-\frac{\cos(\boldsymbol{z}^{\mathcal{R}}_{i,m}, \boldsymbol{z}^{\mathcal{R}}_{i,m'})}{\tau'}\right)\,. \tag{51}$$

In Appendix D.1, we empirically compare the effects of cross-entropy and Gaussian uniformity.

D Ablation Studies and Hyperparameters

D.1 Effects of Temperature and Uniformity Variant

Table 2: Comparison of different uniformity-based models (uni-gauss and uni-xent with different local temperatures τ') on the evaluation tasks. For each model we show on how many of the 18 evaluation tasks it i) outperformed uniformity-based models with the same type of uniformity loss but with different temperatures τ', ii) other uniformity-based models, iii) LoVT without local losses, iv) LoVT, v) CLIP, and vi) on how many it is the best of all studied methods (including the image-only and supervised baselines studied by Müller et al. [2022a]). In brackets, we additionally show the number of tasks considering the 95%-confidence interval over five evaluation runs. For detailed results we refer to Appendix E.

	Best τ'	Best Uni	> /wo local	> LoVT	> CLIP	Best model
Uni-Gauss $\tau' = 0.2$	10 (3)	6 (1)	10 (5)	6 (2)	9 (5)	3 (1)
Uni-Gauss $\tau' = 0.5$	7 (6)	5 (3)	11 (9)	7 (6)	10 (6)	3 (2)
Uni-Gauss any τ'	-	11 (5)	12 (11)	8 (6)	13 (9)	3 (2)
Uni-Xent $\tau' = 0.2$	13 (7)	5 (1)	12 (7)	7 (5)	12 (7)	2 (1)
Uni-Xent $\tau' = 0.3$	4 (2)	2 (2)	9 (6)	4 (3)	7 (6)	1 (1)
Uni-Xent any τ'	-	7 (3)	14 (8)	7 (5)	12 (9)	3 (2)
Uni (any)	-	-	15 (12)	9 (6)	14 (10)	9 (4)

D.2 Shared Head and Attention Pooling

Table 3: Effect of sharing the projection heads for local representations ($\boldsymbol{z}^{\mathcal{I}}_{i,k}$ and $\boldsymbol{z}^{\mathcal{R}}_{i,m}$, used in the uniformity losses) with the heads for global representations ($\bar{\boldsymbol{z}}^{\mathcal{I}}_{i}$ and $\bar{\boldsymbol{z}}^{\mathcal{R}}_{i}$), and the effect of using attention pooling to compute global representations. Note that attention heads are only shared between representations of the same modality, i.e. image and report representations are always projected independently. When no attention pooling is used, global representations are instead computed using global average pooling. We found that using shared attention heads is in general beneficial. We assume that this assures a stronger coupling between the alignment effect of the global contrastive loss and the distribution priors of the local uniformity losses. We also found that attention pooling is beneficial, confirming the results of Müller et al. [2022b].

Uniformity	τ'	Shared head	Att. Pool	RSNA YOLOv3 Frozen 10%	RSNA Lin. Seg. 10%
Gauss-Uniformity	0.2	✓	✓	**18.4**	48.5
		✓	✗	14.9	49.3
		✗	✓	15.6	48.2
Gauss-Uniformity	0.5	✓	✓	**17.6**	49.4
		✓	✗	15.9	49.2
		✗	✓	16.4	49.1
Xent-Uniformity	0.2	✓	✓	**17.4**	49.3
		✓	✗	16.3	49.2
		✗	✓	16.4	49.2
Xent-Uniformity	0.3	✓	✓	**17.6**	49.2
		✓	✗	15.9	48.5
		✗	✓	17.1	49.6

D.3 Hyperparameter Tuning Results

Table 4: Hyperparameter tuning results of the local temperature τ', tuned using the evaluation task *RSNA YOLOv3 Frozen 10%*. The coefficient η was fixed to 0.25 for Gauss-Unifomity and to 0.5 for Xent-Uniformity. We also show the results on *RSNA Lin. Seg. 10%*. In our main studies we use the two best temperatures per variant (marked in blue).

Uniformity	τ'	RSNA YOLOv3 Frozen 10%	RSNA Lin. Seg. 10%
Gauss-Uniformity	0.1	15.3	48.9
	0.2	**18.4**	48.5
	0.3	16.0	49.3
	0.5	17.6	49.4
	1.0	15.9	49.6
Xent-Uniformity	0.05	16.1	49.1
	0.1	17.2	49.4
	0.2	17.4	49.3
	0.3	**17.6**	49.2
	0.5	14.9	49.7

Table 5: Hyperparameter tuning results of the uniformity loss coefficient η, tuned using the evaluation task *RSNA YOLOv3 Frozen 10%*. We also show the results on *RSNA Lin. Seg. 10%*. In our main studies we use the configurations marked in blue.

Uniformity	τ'	η	RSNA YOLOv3 Frozen 10%	RSNA Lin. Seg. 10%
Gauss-Uniformity	0.2	0.1	15.5	50.0
		0.25	**18.4**	48.5
		0.5	15.7	49.0
Gauss-Uniformity	0.5	0.1	16.5	49.7
		0.25	**17.6**	49.4
		0.5	16.6	49.2
Xent-Uniformity	0.2	0.25	16.2	49.2
		0.5	**17.4**	49.3
		0.75	15.6	49.6
Xent-Uniformity	0.3	0.25	15.8	50.0
		0.5	**17.6**	49.2
		0.75	13.9	49.4

E Detailed Experiment Results

Table 6: Results on the RSNA pneumonia detection tasks with different training set sizes. All results are averaged over five evaluation runs and the 95%-confidence interval over these runs is shown. The best results per task are underlined, the second best results are dash-underlined and the best results per block are highlighted in bold. Note that the *RSNA YOLOv3 Frozen 10%* task was used for tuning of all methods and may therefore not be representative as methods may overfit on this task.

	RSNA YOLOv3 Finetune mAP (%)			RSNA YOLOv3 Frozen mAP (%)			RSNA Lin. Seg. Dice (%)		
	1%	10%	100%	1%	10%	100%	1%	10%	100%
Random	2.4±0.5	5.1±1.2	14.9±1.7	1.0±0.2	4.0±0.3	8.9±0.9	21.9±1.2	5.3±0.0	5.3±0.0
ImageNet [Russakovsky et al., 2015]	**5.0±0.7**	**12.4±0.8**	**19.0±0.2**	**3.6±1.4**	**8.0±0.1**	**15.7±0.3**	**27.5±0.6**	**38.3±0.0**	**43.3±0.0**
CheXpert [Irvin et al., 2019]	**8.3±0.8**	12.4±1.6	**21.3±0.3**	7.0±1.0	14.8±0.8	18.8±0.4	38.9±0.2	45.5±0.2	48.1±0.0
BYOL [Grill et al., 2020]	7.0±1.0	11.9±1.1	18.8±0.2	9.6±0.2	14.0±1.2	**21.0±0.2**	42.9±0.1	47.8±0.2	50.0±0.0
SimCLR [Chen et al., 2020]	6.7±0.5	**12.9±0.5**	20.4±1.8	7.9±1.0	11.9±0.1	19.9±0.2	43.1±0.0	46.0±0.0	48.2±0.0
PixelPro [Xie et al., 2020]	4.8±0.6	12.6±1.2	19.8±0.4	3.1±0.2	6.4±0.5	13.4±0.3	25.9±0.2	34.6±0.0	39.8±0.1
ConVIRT [Zhang et al., 2020]	7.4±1.3	12.7±1.5	18.3±0.4	**9.8±0.3**	14.8±1.1	8.4±1.1	42.1±0.1	47.1±0.2	50.2±0.0
CLIP [Radford et al., 2021]*	7.2±0.8	12.8±1.2	19.7±0.5	9.3±0.4	16.1±1.1	19.6±1.4	44.3±0.1	48.8±0.1	50.7±0.0
LoVT [Müller et al., 2022b]	7.7±1.0	11.7±0.5	17.2±1.3	8.6±1.5	**17.9±0.4**	18.0±0.1	**46.0±0.0**	**49.4±0.0**	**51.5±0.0**
LoVT /wo local	5.3±0.9	**12.5±1.3**	19.9±0.0	6.8±0.8	17.4±0.9	18.1±0.1	44.6±0.1	48.7±0.0	51.2±0.0
+ Uni-Gauss $\tau' = 0.2$	**7.3±0.9**	11.7±1.1	18.0±0.9	7.4±1.4	**18.4±0.8**	**22.1±0.1**	44.7±0.4	48.8±0.1	50.8±0.0
+ Uni-Gauss $\tau' = 0.5$	6.4±0.8	10.6±0.9	19.5±0.2	**8.3±1.3**	17.6±0.1	19.1±1.0	**45.9±0.0**	**49.4±0.0**	50.8±0.0
+ Uni-Xent $\tau' = 0.2$	5.8±1.2	10.8±1.4	**20.0±1.0**	7.0±1.1	17.4±1.4	21.6±0.5	45.7±0.3	49.3±0.0	51.1±0.0
+ Uni-Xent $\tau' = 0.3$	5.6±1.3	11.2±0.9	18.0±2.7	6.6±1.1	17.6±0.6	19.1±0.1	45.7±0.1	49.2±0.0	**51.3±0.0**
Task Nr.	1	2	3	4	5	6	7	8	9

* Modified to use the same image and text encoders as ConVIRT and LoVT.

Table 7: Results on downstream tasks on the COVID Rural, SIIM Pneumothorax, Object CXR, and NIH CXR datasets. All results are averaged over five evaluation runs and the 95%-confidence interval over these runs is shown. The best results per task are underlined, the second best results are dash-underlined and the best results per block are highlighted in bold.

	COVID Rural			SIIM-ACR Pneumoth.		Object CXR			NIH CXR
	UNet Finetune	UNet Frozen	Linear	UNet Finetune	UNet Frozen	YOLOv3 Finetune	YOLOv3 Frozen	Linear	Linear
	Dice (%)	Dice (%)	Dice (%)	Dice (%)	Dice (%)	fROC (%)	fROC (%)	Dice (%)	Avg Dice (%)
Random	34.0±1.1	32.2±1.8	6.0±0.0	23.2±1.0	23.9±1.6	49.5±1.2	28.4±1.4	6.9±0.0	0.5±0.4
ImageNet [Russakovsky et al., 2015]	**43.9±2.0**	**41.9±1.7**	**32.6±0.7**	**38.5±0.9**	**36.9±0.7**	**62.5±0.4**	**52.7±1.3**	**37.8±0.0**	**2.6±1.6**
CheXpert [Irvin et al., 2019]	43.5±4.9	44.1±3.2	32.1±2.0	38.9±0.9	40.7±0.7	62.2±0.6	46.3±1.9	16.5±7.7	8.7±0.6
BYOL [Grill et al., 2020]	46.2±1.6	47.5±1.6	36.9±1.7	43.1±0.6	42.9±0.3	59.6±1.0	55.7±1.0	32.3±0.1	6.0±0.1
SimCLR [Chen et al., 2020]	44.9±2.9	41.4±3.7	33.0±0.0	42.6±0.4	39.2±0.7	61.9±0.8	54.3±1.0	33.2±0.1	13.3±0.5
PixelPro [Xie et al., 2020]	47.0±3.4	38.5±3.9	26.6±0.4	39.3±0.8	39.1±0.3	**63.1±0.7**	46.3±0.2	29.9±0.2	1.8±0.0
ConVIRT [Zhang et al., 2020]	48.8±2.2	44.2±3.1	45.0±3.0	42.5±1.0	42.5±0.2	62.5±0.1	54.0±0.7	37.7±0.1	11.4±0.8
CLIP [Radford et al., 2021]*	49.3±2.0	46.5±2.3	46.2±0.3	42.8±1.5	42.5±0.6	62.9±0.8	55.5±2.1	**39.0±0.0**	12.5±1.0
LoVT [Müller et al., 2022b]	**49.5±1.3**	**49.2±4.6**	**49.2±0.2**	**43.4±0.7**	**43.1±0.6**	61.0±1.3	**55.8±1.1**	37.6±0.2	**13.4±0.8**
LoVT /wo local	48.2±1.1	48.7±3.7	45.6±0.0	**43.5±0.9**	42.6±0.4	60.4±1.6	54.8±1.1	38.5±0.2	**13.2±0.9**
+ Uni-Gauss $\tau' = 0.2$	**51.4±3.1**	48.7±0.8	**45.7±2.3**	43.2±0.8	41.6±0.4	61.8±1.6	54.5±1.1	39.1±0.0	11.7±1.3
+ Uni-Gauss $\tau' = 0.5$	50.3±1.2	48.3±3.7	44.2±0.7	43.1±1.3	44.3±0.4	**63.4±0.8**	**57.8±0.9**	38.9±0.0	8.9±1.6
+ Uni-Xent $\tau' = 0.2$	50.5±1.1	**51.3±4.8**	44.0±0.0	43.4±0.9	44.0±0.5	63.2±1.3	54.9±1.9	**39.8±0.0**	11.9±0.9
+ Uni-Xent $\tau' = 0.3$	46.2±2.1	48.1±1.4	42.1±0.2	43.1±1.4	**44.8±0.3**	60.7±1.2	52.2±2.4	38.7±0.1	8.4±1.2
Task Nr.	10	11	12	13	14	15	16	17	18

* Modified to use the same image and text encoders as ConVIRT and LoVT.