Learning complete and explainable visual representations from itemized text supervision

Yiwei Lyu Chenhui Zhao Soumyanil Banerjee Shixuan Liu Akshay Rao
Akhil Kondepudi Honglak Lee Todd C. Hollon

University of Michigan
yiweilyu@umich.edu

Abstract

Training vision models with language supervision enables general and transferable representations. However, many visual domains, especially non-object-centric domains such as medical imaging and remote sensing, contain itemized text annotations: multiple text items describing distinct and semantically independent findings within a single image. Such supervision differs from standard multi-caption supervision, where captions are redundant or highly overlapping. Here, we introduce ItemizedCLIP, a framework for learning complete and explainable visual representations from itemized text supervision. ItemizedCLIP employs a cross-attention module to produce text item-conditioned visual embeddings and a set of tailored objectives that jointly enforce item independence (distinct regions for distinct items) and representation completeness (coverage of all items). Across four domains with naturally itemized text supervision (brain MRI, head CT, chest CT, remote sensing) and one additional synthetically itemized dataset, ItemizedCLIP achieves substantial improvements in zero-shot performance and fine-grained interpretability over baselines. The resulting ItemizedCLIP representations are semantically grounded, item-differentiable, complete, and visually interpretable.

1. Introduction

Language supervision has emerged as a powerful paradigm for training vision models that generalize across tasks and domains [27]. By aligning images with textual descriptions through contrastive objectives, models such as CLIP learn representations that transfer broadly to classification, retrieval, and captioning without task-specific labels [27, 38]. Recent extensions have shown that enriching supervision with multiple positive captions per image, or multi-positive text supervision, can further improve robustness and semantic coverage [10, 18, 41]. In this setting, captions often express the same visual content in slightly different words, producing high semantic overlap among captions.

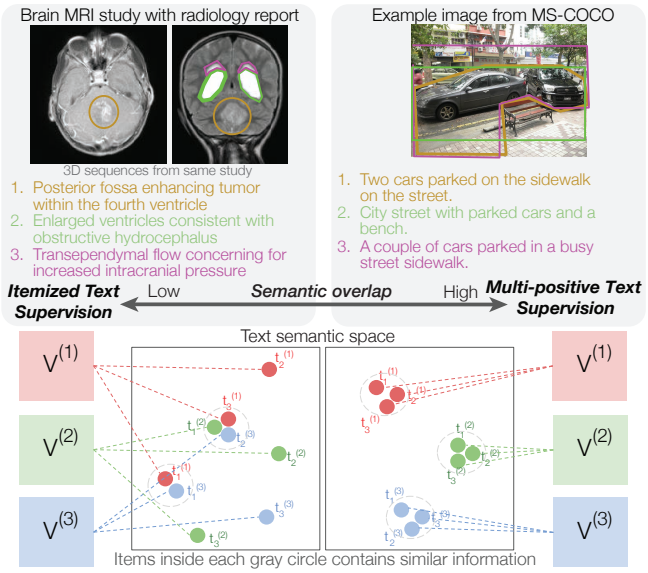


Figure 1. Comparison between itemized text supervision and multi-positive text supervision. (Left) In itemized supervision, each image is paired with multiple independent text items describing distinct findings or regions (e.g., different abnormalities in a brain MRI). These items have low semantic overlap and may recur across unrelated images. Effective visual representations under itemized supervision must capture information from all items rather than any single one. (Right) In multi-positive text supervision, each image is paired with several captions that describe similar visual content with high semantic overlap. (Bottom) Schematic demonstration of the relationship between itemized versus multi-positive settings in the text embedding space. Three visual examples, $V^{(i)}$, each with three associated text items/captions, $t_j^{(i)}$, are shown in the embedding space. Captions cluster together in multi-positive settings. Itemized text do not cluster and can share semantic information across visual examples.

However, many important visual domains do not follow this paradigm. In medical imaging, remote sensing, and

other non-object-centric settings, images are paired not with redundant captions but with itemized textual findings: independent statements that each describe a distinct visual scene within the same image. For instance, a brain MRI report may list separate findings such as "posterior fossa enhancing tumor," "ventricular enlargement," and "transependymal flow." These text items have minimal semantic overlap and refer to spatially disjoint regions. We call this setting **itemized text supervision** (Figure 1).

Itemized text supervision presents a unique learning challenge. Under multi-positive text supervision, a model is rewarded for matching an image with any one of its captions, since they all describe similar content. In contrast, under itemized supervision, a model must represent all text items simultaneously: omitting any one may correspond to missing a critical finding, such as a tumor or hemorrhage in a brain MRI, leading to a misdiagnosis. Effective representations must therefore satisfy two key properties:

- *Item independence*: features associated with different text items should remain distinguishable and localized to their corresponding regions.
- *Representation completeness*: the combined visual embedding should encode information from all items, ensuring full coverage of visual content.

Existing CLIP-style and multi-positive approaches do not satisfy these criteria. For example, multi-positive contrastive objectives explicitly pull together embeddings of all positive captions for the same image, contradicting the independence requirement. Prior models trained for radiology using CLIP-style objectives concatenate all text items into a single caption [14, 24, 40], discarding the compositional structure inherent in itemized supervision.

To address this gap, we propose ItemizedCLIP, a framework for learning complete and explainable visual representations directly from itemized text supervision. ItemizedCLIP employs a cross-attention module to generate item text-conditioned visual embeddings and introduces a set of training objectives specifically designed for this regime. These include an Itemized Local Alignment (ILA) objective that aligns each text item with its corresponding visual evidence with additional mechanisms for completeness and robustness, an Inter-Item Separation (IIS) loss that enforces differentiation between items, a Multi-Positive Siglip (MPS) objective for global alignment, and a Key Token Alignment (KTA) objective to reinforce alignment between each text item and the corresponding small subset of visual tokens. Together, these components allow ItemizedCLIP to learn localized, granular, comprehensive, and interpretable visual features without explicit region annotations.

We evaluate ItemizedCLIP across four domains with naturally itemized text supervision (brain MRI, head CT, chest CT, and remote sensing) and one additional dataset with synthetically itemized captions. Across all five domains, ItemizedCLIP outperforms CLIP-style models and other baselines in zero-shot evaluations. Furthermore, visualization analyses reveal that ItemizedCLIP's attention maps align with human-interpretable findings, demonstrating inherent explainability and grounding. Our main contributions are:

1. We formalize itemized text supervision as a distinct and practically important paradigm for language-supervised learning, and this underexplored paradigm naturally occurs in many application domains.
2. We introduce ItemizedCLIP, a method that jointly enforces item independence and representation completeness through cross-attention and new loss formulations.
3. We provide comprehensive empirical and qualitative evaluations across five domains with itemized text supervision, and ItemizedCLIP consistently achieves strong zero-shot performance, item differentiability, representation completeness and interpretability.

	Llip [19]	Dream-LIP [41]	FLAIR [32]	Itemized-CLIP
Text-conditioned visual representations	✓	✓	✓	✓
Visual token localization from text	×	✓	✓	✓
Zero-shot with local visual tokens	×	×	✓	✓
Enforces item independence	×	✓	×	✓
Enforces representation completeness	×	×	×	✓

Table 1. ItemizedCLIP is designed specifically for itemized text supervision, which differs from multi-positive supervision assumed in previous works, and enforces both item independence and representation completeness.

2. Related Work

We provide a comprehensive discussion of related work in Appendix A. Here we discuss a few existing works designed for multi-positive supervision that are most closely related to ItemizedCLIP: Llip [19] first proposed to address diversity in multiple positive captions via text-conditioned visual representations generated by a cross-attention with text encoding as key and generated visual mixture tokens as key and value. DreamLIP [41] proposed a text-conditioned visual representation based on cross attention between subcaption encodings and local visual tokens (instead of a few mixture tokens in Llip) to allow localization of subcaptions to specific local tokens. FLAIR [32] modifies the training objective from DreamLIP with different negative pairs (TCS loss) and combines it with Multi-positive SigLIP. These works have made significant progress in learning explainable, fine-grained and localized visual representations from multiple positive long captions, but are not designed for itemized text supervision. ItemizedCLIP adapts FLAIR's TCS+MPS objective and added additional objectives to learn visual representations from itemized text supervision that are not only high-quality, explainable and lo-

calizable, but also satisfy the requirements of itemized text supervision (item differentiability and representation completeness). We summarize the comparison between these works in Table 1.

3. ItemizedCLIP Methodology

3.1. Architecture Overview

Like most CLIP-based frameworks, ItemizedCLIP consists of a visual encoder E_V (vision transformer with CLS token) and a text encoder E_T. ItemizedCLIP also contains an additional multi-headed attention `CrossAttn` with linear qkv projection layers. The training data is a paired visual-language $D = \{V^{(i)}, T^{(i)}\}_{i=1}^{N}$ under itemized text supervision, where each visual input $V^{(i)}$ can be patched into m patches $V^{(i)} = \{p_1^{(i)}, p_2^{(i)}, ..., p_m^{(i)}\}$ and each text input $T^{(i)}$ contains n_i text items $T_i = \{item_1^{(i)}, item_2^{(i)}, ..., item_{n_i}^{(i)}\}$. n_i may vary across different $T^{(i)}$.

During the forward pass, given $(V^{(i)}, T^{(i)})$, we encode the text representation $t_j^{(i)}$ for each $item_j^{(i)}$ separately: $t_j^{(i)} = E_T(item_j^{(i)})$, and we obtain two visual representations $(vg^{(i)}, vp^{(i)}) = E_V(V^{(i)})$ where $vg^{(i)}$ is the global visual representation obtained from the CLS token output of E_V, while $vp^{(i)} = \{vp_1^{(i)}, vp_2^{(i)}, ..., vp_m^{(i)}\}$ is the patch-level (i.e. visual-token level) visual representation from the final transformer layer output of E_V over each of the m patches.

3.2. ItemizedCLIP Training Objective

ItemizedCLIP introduces a family of objectives tailored to the characteristics of itemized text supervision (Figure 2).

(a) Itemized Local Alignment (ILA) We begin from Text Conditioned SigLIP (TCS) from FLAIR [32] and adapted it for itemized text supervision. TCS first generates a text-conditioned image representation and computes its cosine similarity with the text item: given text item t and visual patch representations vp, we compute $TCSim(t, vp) = CS(t, \texttt{CrossAttn}(t, vp, vp))$ (where CS is cosine similarity, and `CrossAttn` is the multi-head cross-attention mechanism that uses text item representation t to query visual patch representations vp to compute the text-conditioned visual representation). TCS uses a SigLIP objective to maximize $TCSim(t, vp)$ if the text item belongs to the visual, and minimize $TCSim(t, vp)$ otherwise. In practice, when given an input batch $B = \{V^{(i)}, T^{(i)}\}_{i=1}^{|B|}$, for each i, we compute $TCSim(t_j^{(i)}, vp^{(i)})$ for all $j = 1, 2, ..., n_i$ as positive pairs to be maximized, and we randomly sample one text item representation from each other visual-text pair in the batch ($t_r^{(k)}$, where $k \neq i$ and $r \sim \texttt{random}(1, 2, ..., n_k)$) as negative pairs with $vp^{(i)}$ (i.e. minimizing $TCSim(t_r^{(k)}, vp^{(i)})$) for computational efficiency. So the overall TCS objective for the batch would be:

$$\mathcal{L}_{\text{TCS}}(B) = -\frac{1}{|B|}\sum_{i=1}^{|B|}\left[\sum_{j=1}^{n_i}\text{SigL}\big(TCSim(t_j^{(i)}, vp^{(i)}), 1, b, \tau\big) + \sum_{k\neq i}\text{SigL}\big(TCSim(t_{r_k}^{(k)}, vp^{(i)}), -1, b, \tau\big)\right]$$

where r_k is a random integer between 1 and n_i, SigL represents the single-pair SigLIP objective $\text{SigL}(k, z, b, \tau) = \log(\frac{1}{1+e^{z(-\tau k+b)}})$, b and τ are trainable bias and temperature parameters, and z is the sign variable.

In ItemizedCLIP, we make two modifications to TCS to obtain the ILA objective. To further encourage the model to generate visual representations that captures information from all positive text items (*completeness* criteria), we perform **Upweighting Worst Positive (UWP)**: for each visual $V^{(i)}$, we upweight the positive pair loss between $vp^{(i)}$ and its corresponding text item with the lowest $TCSim$ by a factor of w_{uwp}, which is a hyperparameter. In addition, we perform masked attention within $TCSim$ calculation during training with a randomly generated binary mask from Bernoulli(p_{mask}), where p_{mask} is a hyperparameter controlling the level of masking. So ILA (i.e. TCS+UWP+masking) is:

$$\mathcal{L}_{\text{ILA}}(B) = -\frac{1}{|B|}\sum_{i=1}^{|B|}\left[\sum_{j=1}^{n_i} w_j^{(i)}\text{SigL}\big(TCSim_{\text{masked}}(t_j^{(i)}, vp^{(i)}, p_{\text{mask}}), 1, b, \tau\big) + \sum_{k\neq i}\text{SigL}\big(TCSim_{\text{masked}}(t_{r_k}^{(k)}, vp^{(i)}, p_{\text{mask}}), -1, b, \tau\big)\right]$$

where $w_j^{(i)} = \begin{cases} w_{\text{uwp}} & j = \text{argmin}_j TCSim(t_j^{(i)}, vp^{(i)}) \\ 1 & \text{otherwise} \end{cases}$.

UWP forces the model to learn more from the positive pairs that it currently thinks is the "least positive", therefore improving representation completeness. The masked attention prevents each attention head within the `CrossAttn` module from seeing a random subset of the visual tokens during training, thereby preventing overfitting and yielding more robust visual grounding.

(b) Inter-Item Separation (IIS) The item-independence property implies that different text items should attend to different regions of the image, so we introduce an additional inter-item loss to encourage this. IIS works by treating the text-conditioned visual representation $v_{i,j}^{tc} = \texttt{CrossAttn}(t_j^{(i)}, vp^{(i)}, vp^{(i)})$ and a text item from the same visual $t_k^{(i)}$ as negative pairs if $k \neq j$ and as positive pairs if $k = j$. Formally, we have

$$\mathcal{L}_{\text{IIS}}(B) = -\frac{1}{|B|}\sum_{i=1}^{|B|}\sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\text{SigL}(CS(t_k^{(i)}, v_{i,j}^{tc}), \mathcal{I}_{k=j}, b, \tau)$$

where $\mathcal{I}_{k=j} = 1$ if $k = j$ and is -1 otherwise. During training, we reuse the masked cross-attention results from

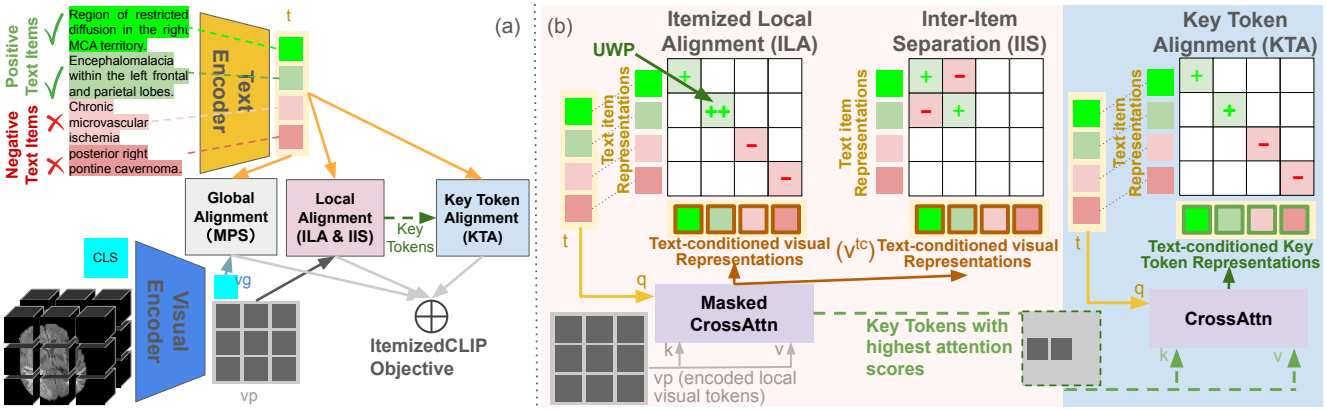


Figure 2. Overview of ItemizedCLIP. **(a)** ItemizedCLIP includes several SigLIP-based objective components: Itemized Local Alignment **(ILA)**, Inter-Item Separation **(IIS)**, Multi-positive SigLIP **(MPS)** and Key Token Alignment **(KTA)**, and the overall objective is their weighted sum. MPS follows [32] and applies a SigLIP objective between the global visual token and each text item representation. **(b)** ILA, IIS and KTA losses are all based on cosine similarities between text item encodings and cross-attention output between text encodings and encoded visual tokens. The grids show which cosine similarities are being used as positive (+) / negative (-) pairs for each objective. ILA mainly aims to distinguish positive text items from negative ones, while IIS aims to distinguish between positive text items. Upweighting Worst Positive **(UWP)** upweights (++) the current weakest positive pair in ILA when computing the ILA loss. The key visual tokens from ILA's masked cross attention with the highest attention scores are used for KTA, which computes an objective similar to ILA but with key tokens only and no masking.

ILA. IIS loss enhances text item differentiability by effectively pushing apart $v^{tc}_{i,j}$ and $v^{tc}_{i,k}$ for all $j \neq k$, thereby forcing the cross-attention module to attend to distinct parts of the visual representation for different text items. FLAIR [32] explored an IIS-like objective in its Appendix but abandoned it as it didn't work well under multi-positive supervision. However, we found that the IIS objective is very effective for itemized text supervision (see section 5.3).

(c) Multi-positive SigLIP (MPS) Multi-positive SigLIP loss is the SigLIP loss defined under multi-positive supervision, where the global visual representation $vg^{(i)}$:

$$\mathcal{L}_{\text{MPS}}(B) = -\frac{1}{|B|}\sum_{i=1}^{|B|}\left[\sum_{j=1}^{n_i}\text{SigL}(CS(t_j^{(i)}, vg^{(i)}), 1, b, \tau) + \sum_{k\neq i}\text{SigL}\big(CS(t_{r_k}^{(k)}, vg^{(i)}), -1, b, \tau\big)\right]$$

We follow [32] and only include one random text item from every other visual in the batch as negative pairs. Although MPS has the effect of pushing representations of different text items paired to the same visual (i.e. $t_j^{(i)}$ with different j) closer to each other, which is undesirable in itemized text supervision due to item independence, we empirically found that adding lower-weighted MPS loss help improve the model's awareness of global visual properties which benefits downstream zero-shot classifications.

(d) Key Token Alignment (KTA) KTA loss guides the model to align text items with compact, high-attention visual tokens. For each text item representation t and visual tokens vp, we obtain the attention map from `CrossAttn(t,vp,vp)` (mean-aggregated across all attention heads) and define $KT(t, vp)$ to be the subset of vp that has the top $K\%$ attention scores in the attention map, where K is a hyperparameter. To strengthen localization, we perform TCS loss with only the key tokens ($KT(t, vp)$ instead of vp):

$$\mathcal{L}_{\text{KTA}}(B) = -\frac{1}{|B|}\sum_{i=1}^{|B|}\left[\sum_{j=1}^{n_i}\text{SigL}\big(TCSim(t_j^{(i)}, KT(t_j^{(i)}, vp^{(i)})), 1, b, \tau\big) + \sum_{k\neq i}\text{SigL}\big(TCSim(t_{r_k}^{(k)}, KT(t_{r_k}^{(k)}, vp^{(i)})), -1, b, \tau\big)\right]$$

(e) Overall loss of ItemizedCLIP is the weighted sum of the four objectives ((a)-(d)):

$$\mathcal{L}_{\text{all}} = \mathcal{L}_{\text{ILA}} + \lambda_{\text{IIS}}\mathcal{L}_{\text{IIS}} + \lambda_{\text{MPS}}\mathcal{L}_{\text{MPS}} + \lambda_{\text{KTA}}\mathcal{L}_{\text{KTA}}$$

3.3. Zero-shot inference

When performing zero-shot inference with ItemizedCLIP, given a visual $V^{(i)}$ and a set of classes in text description $c_1, c_2, ...$, we first encode them with the visual/text encoders to obtain visual tokens $vp^{(i)}$ and text representations $t_{c_1}, t_{c_2}, ...$, and then our prediction logit for class k is simply $TCSim(t_{c_k}, vp^{(i)})$.

4. Experiments

We comprehensively evaluate ItemizedCLIP against SOTA baselines in 4 domains with naturally itemized text supervision: whole-study brain MRI, whole-study head CT, chest CT, and remote sensing. In addition, we create a dataset

Categories	Cyst	Develo-pmental	Infec-tious	Inflam-matory	Sellar	Spine	Structural	Surgical	Trauma	Tumor	Vascular	Overall mAUC
# tasks in category	2	3	2	2	3	1	8	3	3	13	12	52
Prima [24]	64.4	83.2	69.7	81.6	79.9	85.7	68.7	75.2	69.0	80.3	75.4	75.7
FullViT [40]	76.1	89.6	76.3	93.4	88.7	81.7	76.8	88.0	75.4	88.3	78.5	82.7
HLIP [40]	77.8	87.2	77.2	**91.2**	88.6	86.1	80.7	85.6	81.3	89.7	77.6	83.7
ItemizedCLIP (Ours)	**84.6**	**96.8**	**88.4**	87.0	**91.2**	**94.3**	**88.7**	**95.3**	**85.9**	**92.5**	**89.1**	**90.5**

Table 2. Results for zero-shot evaluation on UM220K prospective test set [24]. The numbers reported are mean AUC across tasks within each of the 11 categories, as well as across all 52 tasks. Individual task AUC is reported in Figure 13 in Appendix.

	Pub-Brain-5-gt						Pub-Brain-5						12-metric
	Str-oke	Gli-oma	Menin-gioma	Meta-stasis	Tumor (3-way)	Disease (5-way)	Str-oke	Gli-oma	Menin-gioma	Meta-stasis	Tumor (3-way)	Disease (5-way)	Mean BAcc
BiomedCLIP [39]	66.7	88.2	63.6	59.8	50.4	31.5	64.7	87.8	63.6	59.8	50.4	31.5	59.8
+annotation	86.4	94.1	75.8	75.0	45.7	45.3	-	-	-	-	-	-	-
ConceptCLIP [26]	69.6	92.1	57.8	69.5	35.2	31.6	66.8	91.9	57.7	67.9	35.7	30.9	58.9
+annotation	93.6	**97.8**	70.8	76.8	39.4	50.8	-	-	-	-	-	-	-
Prima [24]	61.2	81.0	87.7	53.4	45.9	31.4	78.8	89.3	70.8	64.7	42.8	31.6	61.6
FullViT [40]	76.7	93.5	58.2	58.2	42.1	43.5	72.8	93.4	72.9	63.1	45.7	43.4	63.6
HLIP [40]	95.0	89.2	79.6	73.4	54.8	61.3	91.5	89.2	79.2	78.1	**63.3**	63.9	76.5
ItemizedCLIP (Ours)	**99.0**	96.9	**89.3**	**94.8**	**58.2**	**68.2**	**97.3**	**96.8**	**89.4**	**88.8**	57.2	**67.6**	**83.6**

Table 3. Results for zero-shot evaluation on Pub-Brain-5 and its subset Pub-Brain-5-gt with 2D slice annotations. We follow [40] to report balanced accuracy on each of the 4 diseases, as well as 3-way tumor classification and 5-way disease classification (normal + 4 diseases). All numbers except ItemizedCLIP was reported from [40]. The 2D slice annotations were only used for selecting the relevant 2D slices for BiomedCLIP and ConceptCLIP zero-shot (in the "+annotation" rows), and are not used for any other rows.

with synthetic itemized text supervision for natural images, and evaluate ItemizedCLIP against the same baselines.

4.1. Whole-study Brain MRI

Brain MRI is one of the most important medical modalities for diagnosing brain diseases. A brain MRI contains multiple 3D volumes, or sequences, as part of a single medical imaging study. UM220K [24] is the largest dataset of whole-study brain MRI with paired itemized radiology reports, with over 220K study-report pairs (including 3.62M 3D sequences). We train ItemizedCLIP on UM220K with similar architecture and preprocessing as HLIP [40]. See Appendix C.1 for full implementation and baseline details. Then, we perform zero-shot evaluation on UM220K prospective test set [24] (30K temporally separated whole-studies acquired after the training data, with annotations for 52 diagnostic tasks) and Pub-Brain-5 [40] (we follow [40] and report performance on 12 metrics across Pub-Brain-5 and its slice-annotated subset, Pub-Brain-5-gt) and we compare ItemizedCLIP's performance to current SOTA whole-study brain MRI models like Prima [24] and HLIP [40], which are also trained on UM220K. Our zero-shot prompts are identical to Prima and HLIP for fairness of comparison.

We show the results for prospective test set in Table 2 and for Pub-Brain-5 in Table 3 against SOTA baselines. ItemizedCLIP outperformed baselines by large margins overall: +6.8% zero-shot mean AUC across 52 tasks on the prospective test set and +7.1% zero-shot mean balanced accuracy on Pub-Brain-5 compared to previous SOTA, and up to +18% gains on individual tasks.

4.2. Whole-study Head CT

Head CTs also contain several 3D CT sequences in each study. Each 3D CT sequence is presented in 3 different windowing levels: brain, blood, and bone. One of the largest whole-study head CT datasets with paired itemized radiology reports is HeadCT240K [40], which contains over 240K study-report pairs. We use ItemizedCLIP to train a model on this dataset and evaluate on downstream classification tasks against baselines. Our implementation of ItemizedCLIP was identical to HLIP in preprocessing and visual backbone design. We include full implementation details of ItemizedCLIP in Appendix C.2. We perform zero-shot evaluation on the following downstream datasets: (1) prospective test set of HeadCT240K [40] with >21K head CT studies with annotations on 83 different diagnostic tasks; (2) RSNA [11], with >10K head CT studies (>25K sequences) with annotations for 5 diagnostic tasks about intracranial hemorrhage; and (3) CQ500 [6], with ∼500 head CT studies with annotations on 10 diagnostic tasks.

We present the prospective test set results in Table 4 and RSNA/CQ500 results in Table 5 against SOTA baselines. ItemizedCLIP again outperformed zero-shot baselines by a large margin: +9.3% in mean AUC across 83 tasks in the prospective test set and over 5% gains on average over all RSNA/CQ500 tasks, compared to HLIP. RSNA and CQ500 are widely used public benchmarks for evaluating head CT models, so we also compared our performance to recent head CT foundation models, including FM-HeadCT [42], Google-CT [34] and Merlin [2]. Only linear probing performance was available from these foundation models, but

ItemizedCLIP was able to outperform their linear probing performances with zero-shot.

4.3. Single-sequence Chest CT

Another domain where medical images are paired with itemized text descriptions is Chest CT. We follow [40] to train ItemizedCLIP on CT-Rate [14] dataset, which contains over 25K single-sequence Chest CT scans with paired radiologist reports. Each report has an LLM-summarized version that is itemized. We implement ItemizedCLIP using similar data preprocessing steps and visual backbone architecture from current SOTA, HLIP [40] (see Appendix C.3 for implementation details), and we perform the same evaluations as [40]: zero-shot classification on test split of CT-Rate [14] with 16 diagnostic tasks as well as an external dataset, Rad-ChestCT [9], with 14 diagnostic tasks. The results are shown in Table 6. ItemizedCLIP again outperforms all baselines across all metrics.

4.4. Remote Sensing

Itemized text supervision naturally occurs in remote sensing datasets such as RSICD [23], which contains over 10K satellite remote sensing images of various geospatial locations, each coupled with itemized text descriptions. We compare ItemizedCLIP against several CLIP-style baselines on a 30-way classification task. Since RSICD is a relatively small dataset, all compared models start from random initialization without any pre-trained weights and only trained on RSICD training split for fairness of comparison. We include further details about implementation, baselines, and preprocessing in Appendix C.4. We also found that applying Diverse Sampling (DS) [32] improves performance for RSICD (but not for medical tasks, see Appendix D.5), so DS is applied to the text items when training ItemizedCLIP and all baselines (except Vanilla CLIP and SigLIP). We show results in Table 7, where ItemizedCLIP outperformed baselines in most metrics, demonstrating that ItemizedCLIP generalizes well to itemized text supervision outside of medical imaging.

4.5. Proof of concept experiment: synthetic itemized text supervision for natural images

In addition to the domains with naturally itemized text supervision, we also aimed to evaluate if ItemizedCLIP works well for learning natural image representations. We create a synthetic set of itemized captions based on the first 10% of data in CC3M-ReCap [41] called Itemized-cc0.3M, by using LLMs to generate itemized captions from the original captions (see Appendix B for curation details). The scale of the synthetic data roughly matches the other tested datasets (brain MRI and CT, on the order of 10^5 data pairs). For fairness of evaluation, we train all models with fully random initialization on Itemized-cc0.3M data only, with DS enabled, and we evaluate on MS-COCO [20] and Flickr [35] retrieval. The results are shown in Table 8, and ItemizedCLIP outperformed all baselines. Implementation and evaluation details are in Appendix C.5.

Note that the synthetic itemized text supervision is significantly more challenging than the original multi-positive supervision, as the total amount of words per image available for training is over 10X less in Itemized-cc0.3M compared to CC3M-Recap. The goal of this experiment is to serve as a proof-of-concept that ItemizedCLIP is a better option compared to existing methods when only itemized text supervision is available.

5. Analysis

5.1. Item-level explainability

ItemizedCLIP-trained models are naturally explainable through attention scores from `CrossAttn` module, similar to [32]. Text item cross-attention provides a direct method to evaluate if ItemizedCLIP's representations are grounded. Given encoded visual tokens vp and text item t, ItemizedCLIP's explanations can be generated by taking the attention scores from $\texttt{CrossAttn}(t, vp, vp)$ and averaged across all attention heads. We include some visualization examples in Figure 3, with additional examples in Appendix E.1. Explainability and grounding are especially essential in medical settings. One key advantage of ItemizedCLIP is that, while existing works like Prima [24] and HLIP [40] can only explain classification logits through LIME [28] or attention map visualization, ItemizedCLIP can generate a visualization from natural language input, which enables verification of the model's understanding of fine-grained details. For example, in Figure 4(a), ItemizedCLIP visualizes 3 different positive text items correctly on the same brain MRI. In Figure 4(b), ItemizedCLIP generates visualizations that correctly point to two separate locations of the same pathology, confirming ItemizedCLIP's accurate understanding of brain anatomy without explicit supervision from segmentation masks.

5.2. Region-based text item retrieval

Under itemized text supervision, each text item often only refers to a specific small region within the image. We found that ItemizedCLIP has the capability to retrieve relevant text items given just a specific region of the image. Given encoded visual tokens vp of the entire visual input, we take the subset of tokens vp' within the region-of-interest, and retrieve text items t from the text corpus that have the highest $TCSim(t, vp')$. We show an example in Figure 5(a) and additional examples in Appendix E.2.

	Cystic	Vascular	Trauma	Structural	Surgical	Tumor	Degenerative	ENT	Infectious	Congenital	Orbital	Overall mAUC
# tasks in category	4	18	11	7	11	10	2	12	3	2	3	83
FullViT	68.0	76.7	75.3	75.5	75.5	67.7	69.2	69.5	73.5	79.6	64.3	73.0
HLIP	67.5	77.0	78.0	78.7	75.3	73.8	71.3	74.6	72.3	**81.4**	80.3	75.8
ItemizedCLIP (Ours)	**71.8**	**85.7**	**87.2**	**87.1**	**89.7**	**78.4**	**81.1**	**86.8**	**84.5**	80.1	**91.8**	**85.1**

Table 4. Results for zero-shot evaluation HeadCT240K's prospective test set [40]. The numbers reported are mean AUC across tasks within each of the 11 categories, as well as across all 83 tasks. Individual task AUC is reported in Figure 14 in Appendix.

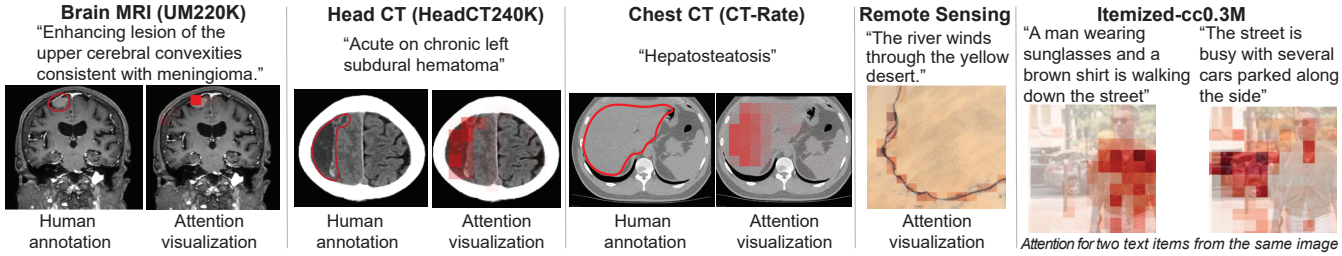


Figure 3. Examples of `CrossAttn` visualization of text items from each of the 5 domains. On the 3D medical imaging domains, we only show attention on one slice from one sequence here due to space, but the attention visualization spans across all slices and sequences within each study, and we show expanded examples in Figure 16. The human annotations were created by a medical professional to illustrate the ground truth region of interest with respect to the text item.

Models	Inference Type	RSNA 5-tasks mAUC	CQ500 10-tasks mAUC
FullViT [24]	Zero Shot	83.3	73.1
HLIP [24]	Zero Shot	85.7	83.1
ItemizedCLIP (Ours)	Zero Shot	**91.5**	**90.0**
FM-HeadCT [42]	Linear Probing	91.3	80.0
Google-CT [34]	Linear Probing	87.2	76.1
Merlin [2]	Linear Probing	73.6	55.1

Table 5. Results for zero-shot evaluation on head CT public benchmarks RSNA [11] and CQ500 [6]. We report mean AUC across all tasks within each dataset. Per-task results are included in Table 13 and Table 14 in Appendix D.3. We also include **linear probing** performance from three recent head CT foundation models, which were outperformed by ItemizedCLIP **zero-shot**.

	CT-Rate (16 tasks)			Rad-ChestCT (14 tasks)		
	Mean AUC	Mean BAcc	Mean wF1	Mean AUC	Mean BAcc	Mean wF1
CT-CLIP [14]	73.3	66.9	70.8	63.3	59.9	64.7
BIUD [4]	71.3	68.1	71.6	62.9	60.6	65.2
Merlin [2]	72.8	67.2	70.9	64.4	61.9	66.3
fVLM [31]	77.8	71.8	75.1	68.0	64.7	68.8
HLIP-RA [40]	77.7	71.4	74.7	72.3	68.4	72.1
HLIP-SA [40]	78.7	72.4	75.5	71.7	67.7	71.4
ItemizedCLIP	**83.2**	**76.5**	**78.9**	**74.7**	**69.8**	**73.2**

Table 6. Results for zero-shot evaluation on single-sequence chest CT classification tasks, on the test split of CT-Rate [14] as well as on an external validation dataset, Rad-ChestCT [9]. Mean zero-shot AUC, balanced accuracy, and weighted F1 scores are reported for each model, following [40] and [31]. HLIP-RA and HLIP-SA refers to HLIP trained with raw annotation and summarized annotations respectively, as reported in [40]. Per-task performance for HLIP and ItemizedCLIP is in Table 15 in Appendix D.4.

	Mean Rank (out of 30)	Top-1 Acc	Top-5 Acc
Vanilla CLIP [27]	5.08	34.2	74.0
Vanilla SigLIP [38]	4.79	38.2	75.7
DreamLIP [41]	4.99	32.2	67.6
Multi-positive SigLIP [32]	4.76	**46.3**	75.4
FLAIR [32]	4.10	41.8	76.1
ItemizedCLIP (Ours)	**3.86**	46.2	**78.7**

Table 7. Result of zero-shot 30-way classification on remote sensing images (test split of RSICD [23] dataset). All models are randomly initialized and trained only with RSICD data.

	MSCOCO				FLICKR			
	I@1	I@10	T@1	T@10	I@1	I@10	T@1	T@10
Vanilla CLIP [27]	1.3	6.6	2.3	10.4	3.1	11.0	4.7	18.4
Vanilla SigLIP [38]	2.1	9.9	3.5	15.6	4.7	16.5	6.9	25.7
DreamLIP [41]	4.0	18.2	6.0	23.2	9.6	30.0	13.6	39.2
Multi-positive SigLIP [32]	4.0	18.0	6.0	23.3	9.2	30.3	11.9	36.9
FLAIR [32]	5.4	22.5	**8.3**	29.4	11.9	36.1	16.7	46.0
ItemizedCLIP (Ours)	**6.1**	**24.1**	8.1	**31.1**	**14.4**	**38.8**	**19.2**	**51.8**

Table 8. Results for zero-shot evaluation on MSCOCO [20] and Flickr [35]. All models are randomly initialized and trained with Itemized-cc0.3M data only.

5.3. Better item differentiability via IIS

Visual representations from itemized text supervision should be able to differentiate between two positive text items and attend to different visual regions when queried with different positive text items (see Figure 4). IIS is the key objective that helps ItemizedCLIP achieve this. We design a metric to evaluate a model's item differentiability: mean Attention Map Similarity (mAMS), which measures average cosine similarity between visualization attention maps of two different positive items. We show mAMS

	Ablations				Completeness		Segmentation
	Brain MRI Pub-brain-5 12-task mBAcc	Chest CT Rad-ChestCT 14-task mAUC	Remote Sensing RSICD Top-1 Acc	Itemized -cc0.3M Flickr Retrieval T@1	Brain MRI Prospective Set MLL*100	Head CT Prospective Set MLL*100	Brain MRI BraTS2021 Zero-shot Tumor Segmentation mIoU
Equal-weight TCS+MPS (FLAIR [32])	78.7	73.9	41.8	16.7	31.12	41.00	11.4
Text-conditioned SigLIP (TCS)	79.9	73.1	41.6	16.4	38.56	40.91	9.1
+ Inter-Item Separation (IIS)	81.2 $^{+1.3}$	73.7 $^{+0.6}$	42.0 $^{+0.4}$	16.5 $^{+0.1}$	40.29 $^{+1.73}$	42.28 $^{+1.37}$	6.5 $^{-2.6}$
+ Multi-Positive SigLIP (MPS)	81.9 $^{+0.7}$	74.2 $^{+0.5}$	42.8 $^{+0.8}$	16.5 $^{+0.0}$	39.43 $^{-0.86}$	41.71 $^{-1.57}$	16.1 $^{+9.6}$
+ Upweight Worst Positive (UWP)	81.6 $^{-0.3}$	74.1 $^{-0.1}$	42.9 $^{+0.1}$	17.8 $^{+1.3}$	42.96 $^{+2.50}$	43.62 $^{+1.89}$	**18.1** $^{+2.0}$
+ Key Token Alignment (KTA)	82.7 $^{+1.1}$	74.1 $^{+0.0}$	44.6 $^{+1.7}$	17.9 $^{+0.1}$	42.46 $^{-0.50}$	43.29 $^{-0.37}$	15.9 $^{-2.2}$
+ TCS Masking (=ItemizedCLIP)	**83.6** $^{+0.9}$	**74.7** $^{+0.6}$	**46.2** $^{+1.8}$	**19.2** $^{+1.3}$	**43.83** $^{+1.37}$	**43.99** $^{+0.70}$	17.5 $^{+2.6}$

Table 9. Ablation study, representation completeness analysis, and zero-shot tumor segmentation results over incremental inclusion of objective components (ILA is divided into its 3 components: TCS, UWP, and TCS-Masking). In addition, we compare with FLAIR [32] objective (equal-weight TCS+MPS).

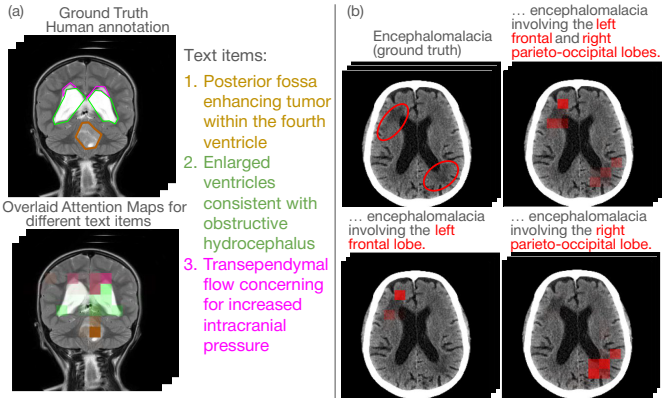


Figure 4. Example of ItemizedCLIP's visualizations of different text items on the same brain MRI/head CT. **(a)** The overlaid attention maps for each of the 3 positive text items align with the ground truth regions. **(b)** A head CT shows encephalomalacia in two regions, and ItemizedCLIP's visualizations can identify both and each one separately based on the text items, indicating ItemizedCLIP's strong grounding and anatomical awareness without explicit anatomical supervision.

of models trained with different objective combinations on brain MRI in Figure 5(b), together with a qualitative example illustrating the difference in Figure 5(c). We show an additional comparative example on brain MRI together with more general examples in Appendix E.3. The results suggest that IIS strongly improves item differentiability.

5.4. Ablation studies

We conduct ablation studies on each of our design choices on metrics from multiple domains. We present the results in "Ablations" columns of Table 9, which shows that each component of ItemizedCLIP improves performance. In addition to the incremental addition of each component, we also show that ItemizedCLIP outperforms FLAIR [32] objective (equal-weight TCS+MPS).

5.5. Representation Completeness

To measure how much ItemizedCLIP complies with the completeness criteria of itemized text supervision, we define a metric called **Mean Lowest Logit (MLL)**: for each visual representation vp with corresponding text items $T = t_1, t_2, ..., t_n$, we calculate lowest logit as $\min_{t \in T} TCSim(t, vp)$, and we average the lowest logit across all images in the testing datasets. A high MLL metric means most visual representations can be matched with all of their corresponding text items, thus satisfying the completeness criteria. We show MLL on brain MRI and head CT in the "Completeness" columns of Table 9, where ItemizedCLIP has the highest MLL scores. Moreover, we observe that UWP, which was designed to improve representation completeness, indeed contributes the most in MLL scores; FLAIR, on the other hand, was designed for multi-positive supervision and thus has lower MLL scores.

5.6. Zero-shot Tumor Segmentation

In addition to zero-shot classification, ItemizedCLIP can perform zero-shot segmentation via the attention heatmap between the encoded text vector of a description of the item to be segmented and encoded visual tokens, following [32]. We perform zero-shot segmentation over BraTS2021 [1], which contains whole-study brain MRIs with ground truth tumor segmentation masks. We report mIoU of the zero-shot masks formed by top 30 visual tokens with the highest attention in the rightmost column of Table 9. The results indicate that both IIS and MPS are needed for strong segmentation performance. Implementation details and example visualizations are in Appendix C.7.

6. Conclusion and Limitations

We present ItemizedCLIP, a language-supervised framework designed to learn complete and explainable visual representations from itemized text supervision, a naturally occurring but underexplored form of supervision where each image is described by multiple independent

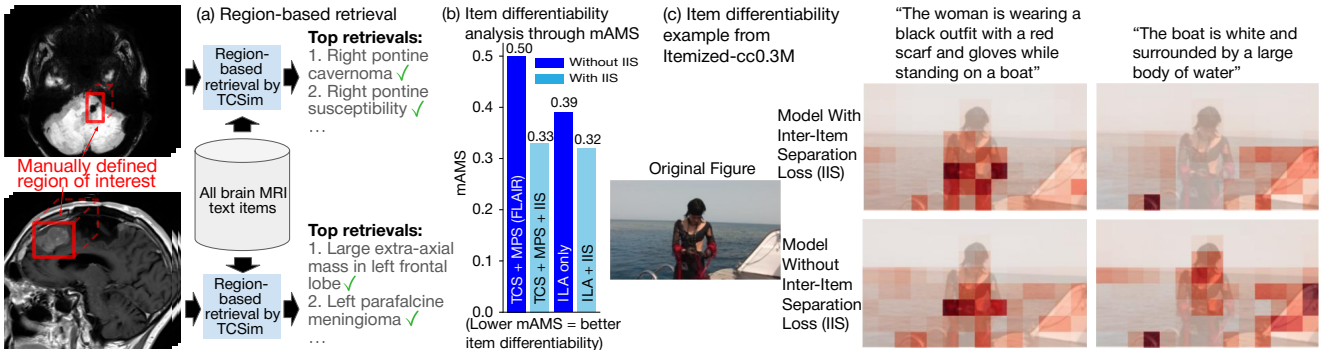


Figure 5. **(a)** Examples of region-based text item retrieval over brain MRIs. We manually provide a region mask. ItemizedCLIP can zero-shot retrieve the text items from the MRI text corpus that best describe the pathology within the region-of-interest (green check). In the bottom example, the retrieved text items describe the meningioma inside the region-of-interest, but not the resection cavity posterior to the region. **(b)** mAMS scores of several training settings with and without IIS on brain MRIs, evaluated on the prospective test set. We can see that when IIS is added to either FLAIR [32] (i.e. TCS+MPS) or ILA, mAMS goes down significantly, indicating that IIS forces the model to attend to different parts of the visuals for different text items. **(c)** An example from Itemized-cc0.3M that demonstrates item differentiability: our model trained with IIS is able to focus on the information mentioned in each text item, while the model trained without IIS has high attention on the woman even when the text item did not mention her.

findings. ItemizedCLIP introduces innovations that enforce item differentiability and representation completeness, yielding interpretable and grounded attention maps and fine-grained alignment between text items and visual evidence. Across four real-world and one synthetic domain, ItemizedCLIP consistently delivers strong zero-shot performance and interpretable item-level grounding, advancing the broader goal of trustworthy vision–language learning.

Despite its generality, our exploration remains bounded by data scale and compute resources. We did not yet examine billion-scale pretraining or the effects of initializing from powerful pretrained vision–language models. Likewise, the approach depends on the quality and granularity of itemized text, which may vary across domains. Future work will explore large-scale fine-tuning, more efficient item-conditioned attention mechanisms, and longitudinal item reasoning across sequential or temporal datasets. We hope these findings inspire broader exploration of itemized text supervision as a natural and informative setting for training interpretable and semantically meaningful vision–language models.

Acknowledgements

References

[1] Ujjwal Baid, Satyam Ghodasara, Suyash Mohan, Michel Bilello, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C Kitamura, Sarthak Pati, et al. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*, 2021. 8

[2] Louis Blankemeier, Joseph Paul Cohen, Ashwin Kumar, Dave Van Veen, Syed Jamal Safdar Gardezi, Magdalini Paschali, Zhihong Chen, Jean-Benoit Delbrouck, Eduardo Reis, Cesar Truyts, et al. Merlin: A vision language foundation model for 3d computed tomography. *arXiv preprint arXiv:2406.06512*, 2024. 5, 7, 8, 9, 10

[3] 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. Making the most of text semantics to improve biomedical vision-language processing, 2022. 6

[4] Weiwei Cao, Jianpeng Zhang, Yingda Xia, Tony CW Mok, Zi Li, Xianghua Ye, Le Lu, Jian Zheng, Yuxing Tang, and Ling Zhang. Bootstrapping chest ct image understanding by distilling knowledge from x-ray expert models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11238–11247, 2024. 7

[5] Souradip Chakraborty, Ekaba Bisong, Shweta Bhatt, Thomas Wagner, Riley Elliott, and Francesco Mosconi. BioMedBERT: A pre-trained biomedical language model for QA and IR. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 669–679, Barcelona, Spain (Online), 2020. International Committee on Computational Linguistics. 3

[6] Sasank Chilamkurthy, Rohit Ghosh, Swetha Tanamala, Mustafa Biviji, Norbert G Campeau, Vasantha Kumar Venugopal, Vidur Mahajan, Pooja Rao, and Prashant Warier. Deep learning algorithms for detection of critical findings in head ct scans: a retrospective study. *The Lancet*, 392(10162): 2388–2396, 2018. 5, 7, 8

[7] Xiaoyi Dong, Jianmin Bao, Yinglin Zheng, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. MaskCLIP: Masked self-distillation advances contrastive language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10995–11005, 2023. 1

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov,

Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020. 3

[9] Rachel Lea Draelos, David Dov, Maciej A Mazurowski, Joseph Y Lo, Ricardo Henao, Geoffrey D Rubin, and Lawrence Carin. Machine-learning-based multiple abnormality prediction with large-scale chest computed tomography volumes. *Medical image analysis*, 67:101857, 2021. 6, 7, 8, 12

[10] Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving clip training with language rewrites. *Advances in Neural Information Processing Systems*, 36:35544–35575, 2023. 1

[11] Adam E Flanders, Luciano M Prevedello, George Shih, Safwan S Halabi, Jayashree Kalpathy-Cramer, Robyn Ball, John T Mongan, Anouk Stein, Felipe C Kitamura, Matthew P Lungren, et al. Construction of a machine learning dataset through collaboration: the rsna 2019 brain ct hemorrhage challenge. *Radiology: Artificial Intelligence*, 2 (3):e190211, 2020. 5, 7, 8

[12] Yuting Gao, Jinfeng Liu, Zihan Xu, Jun Zhang, Ke Li, Rongrong Ji, and Chunhua Shen. Pyramidclip: Hierarchical feature alignment for vision-language model pretraining. *Advances in neural information processing systems*, 35:35959–35970, 2022. 1

[13] Yuting Gao, Jinfeng Liu, Zihan Xu, Tong Wu, Enwei Zhang, Ke Li, Jie Yang, Wei Liu, and Xing Sun. Softclip: Softer cross-modal alignment makes clip stronger. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1860–1868, 2024. 1

[14] Ibrahim Ethem Hamamci, Sezgin Er, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Muhammed Furkan Dasdelen, Omer Faruk Durugol, Bastian Wittmann, Tamaz Amiranashvili, et al. Developing generalist foundation models from a multimodal dataset for 3d computed tomography. *arXiv preprint arXiv:2403.17834*, 2024. 2, 6, 7, 1, 5, 8, 12

[15] Hasan Abed Al Kader Hammoud, Hani Itani, Fabio Pizzati, Philip Torr, Adel Bibi, and Bernard Ghanem. Synthclip: Are we ready for a fully synthetic clip training? *arXiv preprint arXiv:2402.01832*, 2024. 1

[16] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. 3, 6

[17] Cijo Jose, Théo Moutakanni, Dahyun Kang, Federico Baldassarre, Timothée Darcet, Hu Xu, Daniel Li, Marc Szafraniec, Michaël Ramamonjisoa, Maxime Oquab, Oriane Siméoni, Huy V. Vo, Patrick Labatut, and Piotr Bojanowski. Dinov2 meets text: A unified framework for image- and pixel-level vision-language alignment, 2024. 1

[18] Zhengfeng Lai, Haotian Zhang, Bowen Zhang, Wentao Wu, Haoping Bai, Aleksei Timofeev, Xianzhi Du, Zhe Gan, Jiulong Shan, Chen-Nee Chuah, et al. Veclip: Improving clip training via visual-enriched captions. In *European Conference on Computer Vision*, pages 111–127. Springer, 2024. 1

[19] Samuel Lavoie, Polina Kirichenko, Mark Ibrahim, Mahmoud Assran, Andrew Gordon Wilson, Aaron Courville, and Nicolas Ballas. Modeling caption diversity in contrastive vision-language pretraining. *arXiv preprint arXiv:2405.00740*, 2024. 2, 1

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 6, 7, 8

[21] Shixuan Liu, Yiwei Lyu, Honglak Lee, and Todd C Hollon. An empirical study of clip fine-tuning with similarity clusters. In *NeurIPS 2024 Workshop on Fine-Tuning in Modern Machine Learning: Principles and Scalability*, 2024. 1

[22] Yanqing Liu, Kai Wang, Wenqi Shao, Ping Luo, Yu Qiao, Mike Zheng Shou, Kaipeng Zhang, and Yang You. Mllms-augmented visual-language representation learning. *arXiv preprint arXiv:2311.18765*, 2023. 1

[23] Xiaoqiang Lu, Binqiang Wang, Xiangtao Zheng, and Xuelong Li. Exploring models and data for remote sensing image caption generation. *IEEE Transactions on Geoscience and Remote Sensing*, 56(4):2183–2195, 2017. 6, 7

[24] Yiwei Lyu, Samir Harake, Asadur Chowdury, Soumyanil Banerjee, Rachel Gologorsky, Shixuan Liu, Anna-Katharina Meissner, Akshay Rao, Chenhui Zhao, Akhil Kondepudi, Cheng Jiang, Xinhai Hou, Rushikesh S. Joshi, Volker Neuschmelting, Ashok Srinivasan, Dawn Kleindorfer, Brian Athey, Vikas Gulani, Aditya Pandey, Honglak Lee, and Todd Hollon. Learning neuroimaging models from health system-scale data, 2025. 2, 5, 6, 7, 1, 3, 4, 8, 10, 11

[25] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. *arXiv preprint arXiv:2112.12750*, 2021. 1

[26] Yuxiang Nie, Sunan He, Yequan Bie, Yihui Wang, Zhixuan Chen, Shu Yang, and Hao Chen. Concept-clip: Towards trustworthy medical ai via concept-enhanced contrastive language-image pre-training. *arXiv preprint arXiv:2501.15579*, 2025. 5

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 1, 7

[28] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144, 2016. 6

[29] Philipp J Rösch, Norbert Oswald, Michaela Geierhos, and Jindřich Libovickỳ. Enhancing conceptual understanding in multimodal contrastive learning through hard negative samples. *arXiv preprint arXiv:2403.02875*, 2024. 1

[30] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, im-

age alt-text dataset for automatic image captioning. In *Proceedings of ACL*, 2018. 2

[31] Zhongyi Shui, Jianpeng Zhang, Weiwei Cao, Sinuo Wang, Ruizhe Guo, Le Lu, Lin Yang, Xianghua Ye, Tingbo Liang, Qi Zhang, et al. Large-scale and fine-grained vision-language pre-training for enhanced ct image understanding. *arXiv preprint arXiv:2501.14548*, 2025. 7, 1, 5

[32] Rui Xiao, Sanghwan Kim, Mariana-Iuliana Georgescu, Zeynep Akata, and Stephan Alaniz. Flair: Vlm with fine-grained language-informed image representations. 2025. 2, 3, 4, 6, 7, 8, 9, 1, 11, 16

[33] Kaicheng Yang, Jiankang Deng, Xiang An, Jiawei Li, Ziyong Feng, Jia Guo, Jing Yang, and Tongliang Liu. Alip: Adaptive language-image pre-training with synthetic caption. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2922–2931, 2023. 1

[34] Lin Yang, Shawn Xu, Andrew Sellergren, Timo Kohlberger, Yuchen Zhou, Ira Ktena, Atilla Kiraly, Faruk Ahmed, Farhad Hormozdiari, Tiam Jaroensri, et al. Advancing multimodal medical capabilities of gemini. *arXiv preprint arXiv:2405.03162*, 2024. 5, 7, 8, 9, 10

[35] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 6, 7, 8

[36] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 1

[37] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *International Conference on Learning Representations*, 2023. 1

[38] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023. 1, 7

[39] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023. 5

[40] Chenhui Zhao, Yiwei Lyu, Asadur Chowdury, Edward Harake, Akhil Kondepudi, Akshay Rao, Xinhai Hou, Honglak Lee, and Todd Hollon. Towards scalable language-image pre-training for 3d medical imaging, 2025. 2, 5, 6, 7, 1, 3, 4, 8, 9, 10, 11, 12

[41] Kecheng Zheng, Yifei Zhang, Wei Wu, Fan Lu, Shuailei Ma, Xin Jin, Wei Chen, and Yujun Shen. Dreamlip: Language-image pre-training with long captions. In *European Conference on Computer Vision (ECCV)*. Springer, 2024. 1, 2, 6, 7, 3

[42] Weicheng Zhu, Haoxu Huang, Huanze Tang, Rushabh Musthyala, Boyang Yu, Long Chen, Emilio Vega, Thomas O'Donnell, Seena Dehkharghani, Jennifer A Frontera, et al. 3d foundation ai model for generalizable disease detection in head computed tomography. *arXiv preprint arXiv:2502.02779*, 2025. 5, 7, 8, 9, 10

Learning complete and explainable visual representations from itemized text supervision

Supplementary Material

A. Expanded related works

A.1. Visual representation learning with language supervision

CLIP [27] enabled visual representation learning from language supervision at large scale, using a contrastive objective to match image representation with text representation. Since then, many methods have been proposed to improve various aspects of CLIP training: SigLIP [38] proposes an alternative contrastive loss formulation that enables better compute distribution and grants more flexibility in objective design, several approaches aim to combine the strength of self-supervised learning into CLIP objectives [7, 17, 25], and others aim to improve CLIP model's awareness of compositions by making the text in negative pairs more difficult [21, 29, 37].

Recently, there has been an increasing demand for visual representations to capture more fine-grained details, and many works have tackled this through enforcing more fine-grained alignment between visual and text modalities: PyramidCLIP [12] attempts to perform alignment from global level to local level through a hierarchically designed objective, SoftCLIP [13] uses soft assignments to allow many-to-many mappings during training, and CoCa [36] creates an additional caption-generation objective that uses a cross-attention layer to pool the local image tokens for caption generation, thus enforcing a more fine-grained image-to-text alignment.

Another sequence of existing research aim to improve visual-language contrastive learning by collecting more diverse positive captions for each image. LaCLIP [10] and VeCLIP [18] uses LLMs to rewrite existing captions into alternative forms, and more recent advancements in Multimodal large language models (MLLMs) allow generating synthetic captions directly from the images using MLLMs [15, 22, 33, 41]. There has also been explorations in how to best incorporate all of these additional captions into CLIP training most effectively: LaCLIP [10] simply randomly selects one caption each step during regular CLIP training; multi-positive versions of CLIP and SigLIP objectives [32, 41] has been proposed where multiple positive captions can be used within a single step of training; and Llip [19] first proposed to use a cross attention between text and visual representations to allow training with diverse positive captions without forcing representations of different positive captions to be close to each other.

Several recent works aim to combine multi-positive caption training with low-level visual-language alignment to achieve high quality fine-grained visual representations. DreamLIP [41] proposed a text-conditioned visual representation based on cross attention between subcaption encodings and local visual tokens (instead of a few mixture tokens in Llip) to allow localization of subcaptions to specific local tokens. FLAIR [32] modifies the training objective from DreamLIP with different negative pairs (TCS loss) and combines it with Multi-positive SigLIP. These works have made significant progress in learning explainable, fine-grained and localized visual representations from multiple positive long captions, but are not designed for itemized text supervision. ItemizedCLIP adapts FLAIR's TCS+MPS objective (adding UWP and masking to TCS to obtain ILA) and added additional objectives (IIS and KTA) with the aim of learning visual representations from itemized text supervision that are not only high-quality, explainable and localizable, but also satisfy the requirements of itemized text supervision (item differentiability and completeness criteria).

A.2. Visual representation learning in medical imaging with itemized captions

In many medical imaging domains (such as brain MRI, brain CT and chest CT that we explored in this paper), the associated language supervision (i.e. human-written reports) are often naturally well itemized. There exists many works that aim to learn visual representations through CLIP-like objectives using reports as supervision, but they often just concatenate the itemized reports into a single positive caption (e.g. Prima [24] for brain MRI, CT-CLIP [14] for chest CT, and HLIP [40] for all 3 domains) and lose the opportunity to learn fine-grained alignment between each item and local visual tokens. fVLM [31] decomposes chest CT reports into one single caption per organ, and uses pre-obtained segmentation masks to align each anatomical region in the chest CT volume with each organ caption in the reports. Although fVLM achieves an organ-level alignment, it requires availability of ground truth segmentation masks and does not directly achieve visual token-level alignment. ItemizedCLIP aims to fully take advantage of the itemized properties of the reports, and directly achieves alignment between each text item and corresponding local visual tokens, thereby allowing more explainable and localizable representations. ItemizedCLIP also adds additional completeness-based objectives (e.g. UWP) to ensure that the learned objective satisfies the completeness criteria, which is especially important within medical domains, as

failing to pick up any of the abnormalities mentioned in a text item could lead to misdiagnosis.

B. Itemized-cc0.3M curation details

CC3M-Recap [41] is a paired image-caption dataset based on CC3M [30], where in addition to the 1 original caption per image from CC3M, CC3M-Recap includes an additional 6 captions (3 long, 3 short) generated automatically with 3 different large VLMs. Training with CC3M-Recap is a typical multi-positive supervision due to high information overlap across the 7 captions. We aim to create synthetic itemized text supervision from these captions by using GPT-4o to write itemized captions given all 7 captions from CC3M-Recap.

Since most of the captions are generated by VLMs, there are often hallucinatory statements within the automatically generated captions. While having a small hallucinatory part of a long caption in a multi-positive supervision can generally be tolerated, it would be very problematic if the hallucinatory part becomes a standalone text item in our synthetic itemized captions. Therefore, we added a verification step to each piece of information being included in our rewritten itemized captions: we call a piece of information "verified" only if it has occurred in at least 2 of 7 original CC3M-Recap captions (it is unlikely for two different VLMs to make the exact same hallucination), and we only allow verified information in our itemized captions. We want our synthesized captions to have minimum information overlap across items, and together covering all verified information (but not unverified ones that only occurred once among the 7 captions from CC3M-Recap), and we prompted GPT-4o to do so with the following system prompt:

> You are a helpful assistant. You will be provided with a set of 7 captions, 3 long ones and 4 short ones (including a raw one). Your goal is to generate an itemized list of summaries of the captions. Your list of items must satisfy: (1) The list should only include all information that has occurred in at least 2 of the provided captions; and (2) there should be no duplicate information across different entries in the summarized list. Each entry should be a plain sentence.
> Your output should first generate a draft list with at least 5 entries, and then iteratively revised the draft list until the two conditions above are satisfied. Provide reason for each round of revision. The most important things to check during revision is the presence of duplicate information as well as whether the source of each information has occurred in at least 2 of the 7 provided captions. You should list the source captions for each of your list entries in each round of revision, and list the source count. If there is less than 2 sources listed for this entry, you should remove the entry in the next round of revision.
> In addition, list the Primary objects and/or properties that are the focus of each entry. If you find that a part of an entry in the list has duplicate information (about object, attribute, or concept) with another, you should remove the information from this entry during revision, while keeping the remaining unduplicated information in the entry. Each piece of information should only appear once across all entries. In addition, if there exists any two entries that are very similar (i.e. describes the same subject's similar attributes), you should combine the information from those two entries into one. Each list entry should primarily focus on a different set of objects or attributes. Your final entries should contain all remaining details after revision.
> When you are done revising, you must output your final list in the following format: ||finallist|| <final entry 1> | < final entry 2> | ...

We show a concrete example with the original 7 captions in Figure 6, together with GPT-generated itemized captions with and without the two-source verification. As shown in red, the original captions contain quite a few pieces of information that is inaccurate, and two-source verification is able to remove them from the itemized captions.

We perform the above GPT-4o rewriting on the captions of the first 300K images in CC3M-Recap to form the Itemized-cc0.3M dataset. We show a few examples of Itemized-cc0.3M dataset in Figure 7. Note that we do not intend to make any claims about whether itemized captions in Itemized-cc0.3M is superior to those in CC3M-Recap or not. Itemized-cc0.3M is heavily filtered, which means it contains much less hallucinations, but it also contains significantly less overall text compared to the captions in CC3M-Recap. The experiments conducted on Itemized-cc0.3M is only intended to show that, in a hypothetical situation where a natural image dataset with only itemized captions is provided, ItemizedCLIP outperforms alternative options.

C. Implementation Details and Hyperparameters

C.1. Whole-study Brain MRI

Brain MRI is indispensable in modern neurosurgery and neuroradiology. When a patient takes a Brain MRI study, the result is a **whole-study** Brain MRI: a collection of many 3D MRI Sequences. Each 3D sequence consists of many 2D slices, and the 2D slices are "stacked" together along the axis perpendicular to the slice plane to form a dense 3D volume depicting the brain. Each 3D sequence has a weighting type (e.g. T1-weighted, T2-weighted, T1-contrast, FLAIR, SWI, DWI, etc) and an orientation (axial, sagittal, coronal), and each whole-study usually contains many sequences with different types and orientations. Therefore, modeling whole-study brain MRI has been particularly challenging due to the sheer amount of information present in a single study: in UM220K [24], a whole-study

7 captions from CC3M-Recap

raw_caption: how to build an industry for dollars

shortIB_caption: a small cabin being built in the middle of a field

longIB_caption: In the image, there is a small black house with a green roof situated in a grassy area surrounded by trees. The house appears to be under construction or renovation, as there are various tools and materials visible around it, such as a hammer, nails, screws, and wood planks. The presence of these objects indicates that the house is being built or repaired, and the green roof adds a unique and eco-friendly feature to the structure.

shortSV_caption: A wooden building with a green roof is under construction.

longSV_captions: The image captures a tranquil scene in a wooded area. Dominating the frame is a small wooden cabin, its green roof contrasting with the surrounding foliage. The cabin's walls are painted a dark brown, and a window with white shutters punctuates one side. A door, also white, is situated on the opposite side. The cabin is elevated on a wooden platform, providing a vantage point over the verdant landscape. A hose, coiled and ready for use, lies on the ground in front of the cabin, hinting at the possibility of water sources nearby. The cabin is nestled amidst nature, with trees and bushes forming a lush backdrop. The precise location of the cabin is not discernible from the image, but its elevated position and the surrounding greenery suggest a peaceful retreat, possibly in a rural or semi-rural setting. There are no discernible texts or countable objects in the image. The relative positions of the objects are such that the cabin is the central focus, with the hose in the foreground and the trees and bushes in the background. The image does not provide any information that allows for a confident determination of object actions or precise object locations beyond what has been described.

shortLLA_caption: A small wooden building is being constructed with a green roof.

longLLA_caption: The image features a small wooden cabin with a green roof, surrounded by a grassy area. The cabin appears to be under construction, as there are several tools and materials scattered around the scene. A person is standing near the cabin, possibly working on the construction or observing the progress. In addition to the cabin, there are two cars parked in the background, one on the left side and the other on the right side of the image. A bench can be seen in the middle of the scene, and a chair is located closer to the right side of the image.

Hallucinations

Itemized captions without verification by two sources:

- A small wooden cabin with a green roof is under construction in a grassy area surrounded by trees and bushes, with tools and materials scattered around.
- The cabin has dark brown walls, a window and door with white shutters, and is elevated on a platform.
- A hose is coiled in front of the cabin, suggesting possible water sources nearby.
- A person is working near the cabin, with two cars parked in the background and a bench and chair visible in the scene.

Itemized captions with verification by two sources:

- A small wooden cabin with a green roof is under construction
- The cabin is surrounded by a grassy area and trees
- Various tools and materials are visible around the cabin, indicating construction activity

Figure 6. An example image with 7 corresponding captions from CC3M-Recap [41]. As shown in red, the VLM-generated captions contains quite a few pieces of hallucinated information. If we do not instruct GPT to perform two-source verification, some of the hallucinated information will go into the itemized dataset. With two-source verification, although some correct information is filtered out by the verification, it largely prevents any hallucinated information to become a completely wrong text item in the itemized captions.

contains over 1600 2D slices on average. After a patient takes a brain MRI, the whole-study is sent to a trained radiologist for interpretation, and the radiologist will write a report associated with the entire study. The radiology report is typically well itemized, with each sentence describing distinct findings or impressions. The first large-scale study on whole-study brain MRI is Prima [24], which trains a CLIP model between whole-studies and their corresponding reports (summarized by GPT into an itemized list of abnormalities only, included as part of UM220K dataset). We show an illustration of a whole-study as well as an example of summarized report in Figure 8.

The data we use to train our model is UM220K [24], same as Prima [24] and HLIP [40]. It is the largest whole-study brain MRI dataset to the best of our knowledge, and it contains over 220K whole-studies with corresponding summarized radiologist reports. In addition, UM220K also contains an additional prospective test set of 30K whole-studies, separated temporally from the 220K training set by study date. All studies also have binary annotations over 52 diagnoses.

HLIP [40] is a model architecture designed to handle the large amount of data in a single radiology study via efficient hierarchical attention. Each 3D sequence is first reshaped to 48x256x256, then center-cropped to 48x224x224, and then is cut into 8x16x16 tokens. Each token is fed through a 3D CNN to be flattened into a 1D vector, and the sequence of tokens (as 1D vectors) is fed into a ViT-base with CLS token to be encoded. Our implementation of ItemizedCLIP uses the identical model architectures as HLIP for both visual encoder (ViT-base [8] with HLIP hierarchical attention, pre-trained from "vit_base_patch16_224.mae" from OpenCLIP [16]) and text encoder (BiomedBERT [5], from "microsoft/BiomedNLP-BiomedBERT-base-uncased-abstract-fulltext" on Huggingface), with one minor deviation: Prima [24] found that including the names of each 3D sequence (e.g. "AX_T2", etc, available as part of the UM220K) slightly improves model performance, so we replaced the learnable sequence-order embeddings in HLIP visual encoder with the sequence name encoder from Prima (fully trainable) that encodes each sequence's name into an embedding. We call this modified architecture "HLIP-SN". The modification is very light-weight and has negligible impact on computation and memory, but yields a small im-

1. A basketball coach is present on the court
2. The coach is wearing a suit and tie
3. A basketball player is present, wearing a blue jersey
4. There is a crowd of spectators watching the game
5. The scene is set in a basketball game environment

1. The image features a hookah (water pipe) with a blue and red color scheme
2. Oranges are present in the scene, both whole and cut in half
3. The hookah is placed on a table with a reflective surface and a white background

1. The background is yellow in color
2. The phrase "It's time to travel" is prominently featured in the images
3. The motorcycle is depicted in a way that conveys a sense of adventure and travel

Figure 7. Additional caption examples from Itemized-cc0.3M.

	mAUC over 52 tasks
HLIP [40]	83.7
ItemizedCLIP with unmodified HLIP	90.2
ItemizedCLIP with HLIP-SN	**90.5**

Table 10. Zero-shot performance on prospective test set of UM220K. We see that ItemizedCLIP significantly improves model performance over HLIP, whether implemented with unmodified HLIP or HLIP-SN. Implementing ItemizedCLIP with HLIP-SN improves performance slightly.

provement in overall performance. To justify this design choice, we conduct an additional ablation study of implementing ItemizedCLIP with unmodified HLIP, and we show the results in Table 10 over the prospective test set. We can see that ItemizedCLIP still significantly outperforms HLIP even when implemented with identical architecture as HLIP, and using HLIP-SN further improves performance by a small amount.

Another minor adjustment to ItemizedCLIP that we make for brain MRIs is called "normal check": around 10% of studies in UM220K training set have no abnormalities mentioned in the radiologist report, so the GPT-summarized report simply states "Study is unremarkable". We simply mark these studies as "normal" and we do not use text from one normal study and visual from another normal study as negative pairs in ILA, MPS and KTA. Applying normal check improves overall model performance, as shown in Table 11.

We train ItemizedCLIP on UM220K training data for

	mAUC over 52 tasks
ItemizedCLIP without normal check	89.6
ItemizedCLIP with normal check	**90.5**

Table 11. Zero-shot performance on prospective test set of UM220K, comparing ItemizedCLIP with and without normal check. Normal check improves performance.

20 epochs, then we evaluate its zero-shot performance on UM220K prospective test set on 52 tasks. In addition, we also evaluate zero-shot performance of ItemizedCLIP on Pub-brain-5 [40], an evaluation benchmark for brain MRI foundation models formed by combining 5 publicly available datasets. Our zero-shot prompts for the prospective test set is identical to those used in the papers of Prima [24] and HLIP [40], where the best of 7 prompts is reported; and our zero-shot prompts for Pub-brain-5 is also identical to the ones from HLIP [40] codebase. The baseline results presented on these evaluation benchmarks all came from [40].

C.2. Whole-study Head CT

Brain CT is also a very important imaging technology for diagnosing brain diseases, especially hemorrhages and traumatic injuries. Similar to brain MRI, a brain CT whole-study also contains many 3D CT sequences of different types, where each sequences contains many 2D slices that are stacked together to form dense 3D structures. In addition, each 3D CT sequence is presented in 3 different views (brain, blood, bone), each obtained by setting a different window level and window width with respect to the original CT scanner HU and emphasizes different kinds of tissues. We illustrate this structure in Figure 9(a). HeadCT240K [40] is one of the largest datasets for whole-study brain CTs, with 240K CT whole-studies with paired reports. Like brain MRIs, after a patient takes a brain CT, the whole-study is sent to a radiologist for interpretation, and the radiologist will write a report associated with the entire study. The radiologist reports are already quite itemized, with each sentence usually talks about findings from unique regions. HeadCT240K also provides GPT-summarized radiologist reports that filtered out normal comments, so the summarized report is an itemized list of distinct anomalies found. We show an example of summarized report in HeadCT240K in Figure 9(b).

We again follow HLIP [40] and preprocess the data by reshaping each 3D view into 48x256x256, center-cropping to 48x224x224, then cut into 8x16x16 tokens, and flatten with a 3D CNN layer before feeding into a ViT Base (the exact same preprocessing steps as brain MRIs above). The training setup is also identical to what we did for brain MRIs: we again use the HLIP-SN architecture for our model trained with ItemizedCLIPwith the same pre-trained model initialization as HLIP and brain MRIs, and we apply

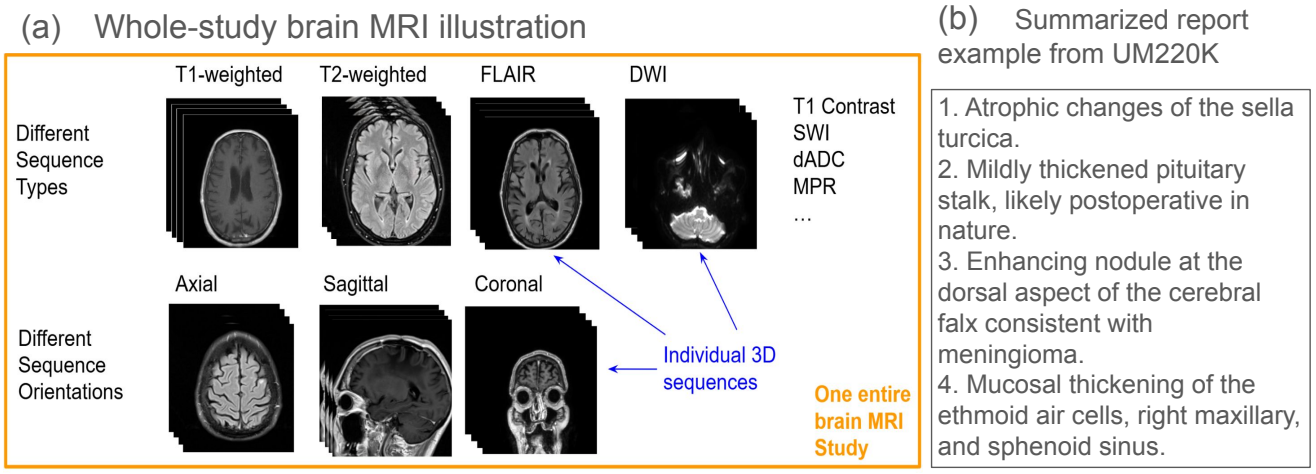


Figure 8. textbf(a): Illustration of brain MRI whole-study. Each brain MRI whole-study contains multiple 3D sequences of different types and orientations. Each 3D sequence contains many 2D slices, which are stacked together to form a dense 3D volume of the brain. **(b)**: An example of summarized report from UM220K. Brain MRIs are a canonical example of itemized text supervision.

normal check (see section C.1 for details). We treat each view of each sequence as a distinct sequence when inputting data into the HLIP encoder, following [40]. The sequence names fed to HLIP-SN for each sequence contains both the name of the sequence and view window information for brain CTs (only sequence names are used for brain MRI, as brain MRI does not have different windowed views). We train ItemizedCLIP on the training set of HeadCT240K.

Like UM220K, HeadCT240K also has a prospective test set that is separated temporally by acquisition date from the training set. The prospective test set contains over 21K brain CT whole-studies with paired summarized reports, as well as diagnostic annotations for 83 diagnoses. We perform zero-shot evaluation on these diagnoses, using the same prompts as [40]. In addition, we perform evaluation on 2 widely used benchmarks for evaluating brain CT foundation models: RSNA [11] has over 10K CT whole-studies (totalling over 25K 3D sequences) and annotations over 6 hemorrhagic diagnoses, and CQ500 [6] has around 500 CT whole-studies and annotations on 10 diagnoses. We follow [40] and [42] to only report performance on 5 out of 6 diagnoses on RSNA. Same prompts were used to evaluate zero-shot classification on RSNA and CQ500 between ItemizedCLIPand baselines (FullViT [40] and HLIP [40]). We include our zero-shot prompts below:

RSNA prompts for 5 diagnoses:
[”Intracranial hemorrhage”, ”Intracranial hemorrhage with intraparenchymal hemorrhage”, ”Intracranial hemorrhage with intraventricular hemorrhage”, ”Intracranial hemorrhage with subarachnoid hemorrhage”, ”Intracranial hemorrhage with subdural hemorrhage”]
CQ500 prompts for 10 diagnoses:
[”Intracranial hemorrhage”, ”Intracranial hemorrhage with intraparenchymal hemorrhage”, ”Intracranial hemorrhage with intraventricular hemorrhage”, ”Intracranial hemorrhage with subdural hemorrhage”, ”Intracranial hemorrhage with epidural hemorrhage”, ”Intracranial hemorrhage with subarachnoid hemorrhage”, ”Bleeding in left side”, ”Bleeding in right side”,”Midline shift”, ”Mass Effect”]

The baseline results reported on all brain CT benchmarks came from [42] and [40].

C.3. Single-sequence Chest CT

The third medical imaging domain we include in our experiments is Single-sequence Chest CT, which contains a single 3D CT sequence taken over a patient's chest (from shoulder to upper abdomen). The dataset we use is CT-Rate [14], which contains 25.6K single-sequence chest CT with paired radiologist reports, and is split into training set and internal validation set (test set with 1.5K sequences). While the raw radiologist reports from CT-Rate can already be considered itemized (as each sentence usually refers to finding or no findings over a specific region of interest), existing works has typically further summarize the reports with LLMs: for example, fVLM [31] used LLM to summarize the reports into findings on 4 organs/regions (lung, heart, esophagus, aorta). We apply a rather simple approach: given a report (itemized by sentences), we simply asked GPT-5 to remove all items that do not describe an abnormality (e.g. items like ”no ... found” or ”no evidence of ...” or ”... appears normal”), similar to the summarized reports for brain MRI and CTs in Prima [24] and HLIP [40]. We show an example chest CT sequence from CT-Rate as well as an example GPT-5 summarized report in Figire 10.

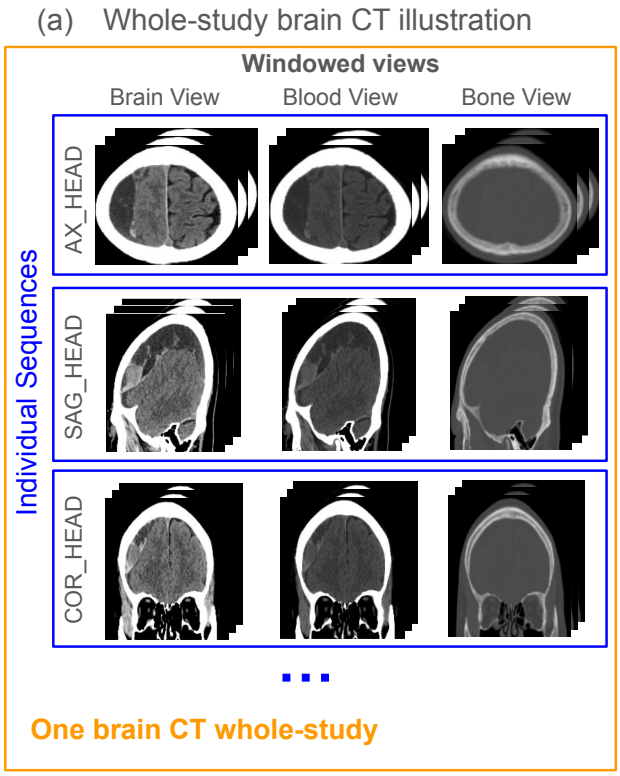


Figure 9. **(a)**: Illustration of brain CT whole-study. Each brain CT whole-study contains multiple 3D sequences of different orientations, and can be viewed in 3 different windows (brain, blood, bone). Each 3D sequence contains many 2D slices, which are stacked together to form a dense 3D volume of the brain. **(b)**: An example of summarized report from HeadCT240K.

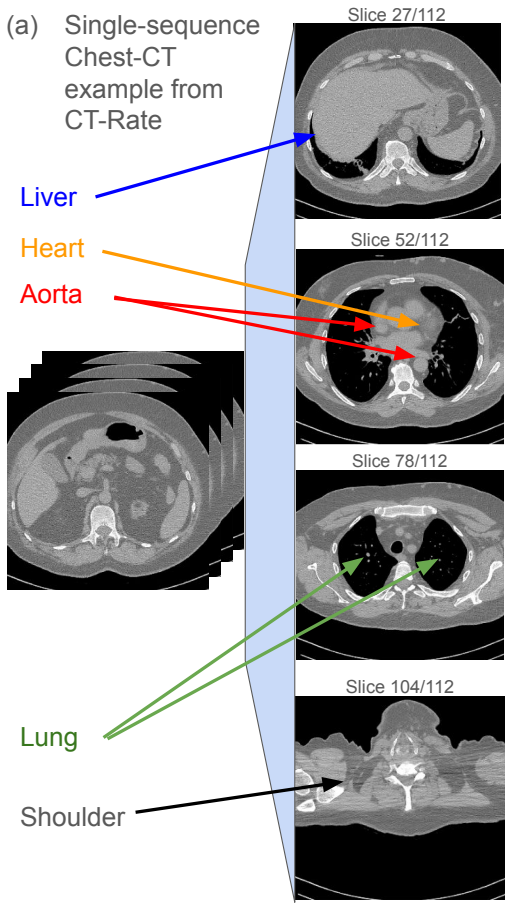


Figure 10. **(a)** An example of single-sequence chest CT from CT-Rate. As shown, it covers many different organs and regions such as shoulder, liver, heart, lung, and aorta. **(b)** An example summarized report for CT-Rate. The summarization simply removes any sentences describing no abnormal findings from the original report.

When training ItemizedCLIP, we again follow the same setup as HLIP [40]: we first reshape each 3D sequence into 112x256x256 and center-crop to 112x224x224, then cut into 8x16x16 tokens which are flattened to 1D vectors through a 3D-CNN layer. The sequence of flattened tokens are then encoded by a ViT-Base model with CLS token. During training, we allow the center-crop to shift randomly by a few pixels as additional data augmentation. We initialize from the same pre-trained model as HLIP [40]: the visual encoder initializes from "vit_base_patch16_224.mae" from OpenCLIP [16]) and text encoder from BiomedVLP-CXR-BERT [3] ("microsoft/BiomedVLP-CXR-BERT-specialized" from Huggingface). We do not freeze the text encoder and allow joint training of visual and text encoders, and we employ normal check.

We follow existing works and evaluate our model on the internal validation split of CT-Rate [14] (1.5K sequences, 16 diagnostic tasks) and an external validation dataset, RadChestCT [9] test split (14 diagnostic tasks, 3.6K). Our zero-shot prompt is as follows (for both datasets, all tasks except "calcification" are in CT-Rate internal validation and all tasks except "mosaic attenuation pattern", "Arterial wall calcification" and "coronary artery wall calcification" are in

Rad-ChestCT):

emphysema: ”findings consistent with emphysema”
atelectasis: ”findings consistent with atelectasis”
lung nodule: ”findings consistent with nodules or nodular density”
lung opacity: ”findings consistent with opacity”
pulmonary fibrotic sequela: ”findings consistent with pulmonary fibrotic sequela”
pleural effusion: ”findings consistent with pleural effusion”
peribronchial thickening: ”findings consistent with peribronchial thickening”
consolidation: ”findings consistent with consolidation”
bronchiectasis: ”findings consistent with bronchiectasis or bronchiectatic changes”
interlobular septal thickening: ”findings consistent with interlobular septal thickening”
cardiomegaly: ”findings consistent with cardiomegaly”
pericardial effusion: ”findings consistent with pericardial effusion”
coronary artery wall calcification: ”findings consistent with coronary artery wall calcification”
hiatal hernia: ”findings consistent with hiatal hernia”
arterial wall calcification: ”findings consistent with arterial wall calcification”
calcification: ”findings consistent with coronary artery wall calcification”

C.4. Remote Sensing

RSICD [23] is a remote sensing dataset with 8.7K satellite image-caption pairs in the training split and 1.1K satellite image-caption pairs in the test split (we follow the data splitting from Huggingface dataset ”arampacha/rsicd”). Each satellite image in RSICD has 5 captions written by different annotators, each usually focuses on one specific object/property of the image. Sometimes there will be duplicates among the 5 captions for a single satellite image, so we use an LLM to deduplicate the captions (only removing duplicated captions, with no change to the wording to any remaining captions) for the training set and the resulting captions are well-itemized with low information overlap across items. We show example satellite images and their paired deduplicated caption items in Figure 11.

We train ItemizedCLIPon the training split of RSICD. The model architecture and initialization is exactly the same as FLAIR [32], with open-clip model config pasted below:

{ ”embed_dim”: 512, ”init_logit_bias”: -10, ”vision_cfg”: { ”image_size”: 224, ”layers”: 12, ”width”: 768, ”patch_size”: 16, ”output_tokens”: true }, ”text_cfg”: { ”context_length”: 100, ”vocab_size”: 49408, ”width”: 512, ”heads”: 8, ”layers”: 12 } }

Figure 11. Examples of remote sensing images with paired deduplicated captions.

We train several CLIP-like baselines on the same training data: Vanilla CLIP [27] and SigLIP [38] are trained with all text items concatenated into one for each image, then apply the standard CLIP or SigLIP loss respectively; Multi-positive SigLIP [32] baseline is trained with MPS loss only; DreamLIP [41] and FLAIR [32] are trained following each paper's original objective setups. We perform diverse sampling (DS) [32] as a data augmentation technique on all methods that takes in multiple positive items per image (DreamLIP, Multi-positive SigLIP, FLAIR, ItemizedCLIP) as it significantly improves performance (see Appendix D.5).

We evaluate each method on the test split of RSICD. There is ground truth annotation of each image's category, out of 30 total categories (e.g. ”airport”, ”dessert”, ”dense residential area”, ”industrial”, ”port”, etc). We perform zero-shot classification on which category each image belongs to out of the 30 categories. We report performance on top-1 accuracy, top-5 accuracy, and mean rank of the correct choice ranked by zero-shot logit of each category. We have 2 zero-shot prompts for each category, and we use the average of the 2 encoded text vectors when performing zero-shot. The 2 zero-shot prompts are as follows:

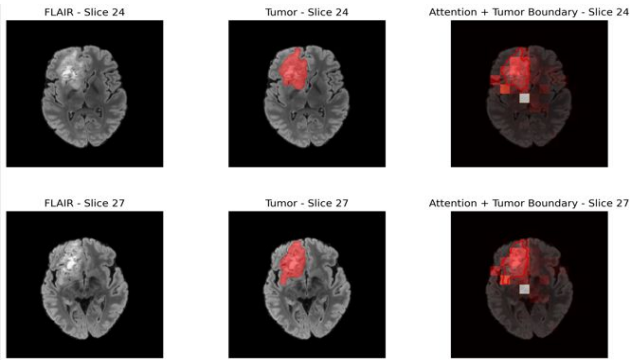


Figure 12. Example zero-shot segmentation result of ItemizedCLIP on BraTS2021 [1]. Two different slices from the same MRI sequence are shown. The middle column shows the ground truth segmentation mask, while the right column shows an overlay of the ground truth with our attention map. Our attention map resembles the ground truth segmentation mask.

"There is <category name>"
"The <category name>"

C.5. Natural Images (Itemized-cc0.3M)

We perform proof-of-concept experiment on Itemized-cc0.3M, a natural-image dataset consisting of 10% of CC3M-Recap images with synthetically itemized text captions (curation process see Appendix B, examples in Figure 7). The setups and training for both ItemizedCLIP and baselines are identical to those described in the previous subsection (for RSICD). We evaluate the models on 2 widely-used retrieval benchmarks: MSCOCO [20] and Flickr [35], and we follow the exact same zero-shot retrieval scripts as FLAIR [32].

C.6. Hyperparameters

We report all hyperparameters used in training ItemizedCLIP on each of the 5 domains in Table 12.

C.7. Zero-shot tumor segmentation

BraTS2021 [1] is a public dataset of curated brain MRI studies, where each study contains a fixed 4 axial sequences (T1, T2, FLAIR, T1CE). All sequences are skull-stripped for privacy preservation. All studies are from patients with Glioma, a type of brain tumor, and each FLAIR sequence has a ground truth segmentation mask. (Note that FLAIR here refers to a sequence type in brain MRI, not the visual-language pre-training method FLAIR [32].) We take our model trained with ItemizedCLIP over UM220K and perform zero-shot segmentation by obtaining the attention map between each whole-study from BraTS2021 and the text prompt "Glioma.", and take the top 30 visual tokens on the FLAIR sequence (out of 6x14x14=1176 total visual tokens for each sequence) as the zero-shot segmentation mask. Then, we compute IoU between zero-shot segmentation mask and the ground truth mask of the glioma, and average over 1251 studies in BraTS2021 to obtain the overall mIoU score reported in Table 9. We show an example from BraTS2021 in Figure 12.

D. Extended Results

D.1. Brain MRI UM220K prospective test set

We show the full task-wise results of 52 diagnostic tasks in UM220K prospective test set in Figure 13. The radar plot style is inspired by [24] and [40]. ItemizedCLIP consistently outperforms baselines across an overwhelming majority of tasks.

D.2. Brain CT prospective test set

We show the full task-wise results of 83 diagnostic tasks in HeadCT240K's prospective test set in Figure 14. ItemizedCLIP consistently outperforms baselines across an overwhelming majority of tasks.

D.3. Brain CT RSNA & CQ500

We show expanded per-task results of brain CT evaluations on RSNA [11] in Table 13 and on CQ500 [6] in Table 14. ItemizedCLIP consistently outperforms baselines in most tasks, and its zero-shot performance even outperforms linear probing performance of several recent CT foundation models (FM-HeadCT [42],Google-CT [34] and Merlin [2]).

D.4. Chest CT CT-Rate & Rad-ChestCT

We show expanded per-task results of chest CT evaluations on CT-Rate [14] and Rad-ChestCT [9] on ItemizedCLIP as well as two top-performing baselines, HLIP-RA and HLIP-SA [40] in Table 15. ItemizedCLIP outperforms baselines in the majority of metrics, often by significant amounts.

D.5. Diverse Sampling analysis on different domains

Diverse sampling (DS) is a data augmentation strategy on the text side developed in [32]. Diverse sampling randomly combines one or more text items together into one single longer text item. We found that applying diverse sampling significantly improves performance for 2D image domains (RSICD and Itemized-cc0.3M) when training with ItemizedCLIP, but it does not work well for medical imaging domains and actually decreases performance of ItemizedCLIP. We show the comparisons in Table 16.

E. Additional qualitative examples

E.1. Text item visualizations

We show additional text item attention visualization examples in Figure 15. In addition, we demonstrate attention

	Brain MRI	Brain CT	Chest CT	Remote Sensing	Itemized-cc0.3M
Weight of IIS loss (λ_{IIS})	1	1	1	1	0.1
Weight of MPS loss (λ_{MPS})	0.01	0.1	0.1	1.5	0.1
KTA loss key token rate (K)	0.05	0.05	0.05	0.2	0.2
Weight of KTA loss (λ_{KTA})	1	1	1	1	0.2
TCS masking rate (p_{mask})	0.1	0.1	0.05	0.4	0.1
Total Batch Size (B)	256	256	512	1024	1024
UWP upweighting factor (w_{uwp})	1.5	1.5	1.5	1.5	2
Maximum number of text items per visual	7	7	10	6	7
Learning Rate	0.000175	0.000175	0.0001	0.0003	0.0005
Weight Decay	0.2	0.5	0.5	1.5	0.8
Optimizer	AdamW	AdamW	AdamW	AdamW	AdamW
Learning Rate Scheduler	Cosine	Cosine	Cosine	Cosine	Cosine
Scheduler beta1	0.9	0.9	0.9	0.9	0.9
Scheduler beta2	0.98	0.98	0.98	0.98	0.98
Total Epochs trained	24	21	80	120	60
Warmup Steps	2000	2000	100	100	2000
Initial SigLIP temperature (τ)	2.659	2.659	2.659	2.659	2.659
Initial SigLIP bias (b)	-10	-10	-10	-10	-10
Diverse Sampling Max Merged Num	-	-	-	3	3
Diverse Sampling Flag Probability	-	-	-	0.5	0.5
`CrossAttn` number of heads	8	8	8	8	8

Table 12. Hyperparameters for training ItemizedCLIP on each domain.

Models	Inference Type	ICH any	ICH intraparenchymal	ICH intraventricular	ICH subarachnoid	ICH subdural	Mean AUC
FullViT [40]	Zero Shot	79.0	82.6	92.7	82.8	79.2	83.3
HLIP [40]	Zero Shot	81.5	88.2	91.4	84.1	83.4	85.7
ItemizedCLIP (Ours)	Zero Shot	<u>92.2</u>	**91.2**	**96.4**	**90.2**	<u>87.4</u>	**91.5**
FM-HeadCT [42]	Linear Probing	**92.6**	<u>90.7</u>	<u>95.6</u>	<u>89.5</u>	**87.9**	91.3
Google-CT [34]	Linear Probing	89.2	88.3	93.2	84.4	81.0	87.2
Merlin [2]	Linear Probing	75.4	75.0	78.6	72.7	66.4	73.6

Table 13. RSNA per-task AUC performance.

visualization over all slices across and within 3D sequences of a brain MRI whole-study in Figure 16. These examples show that, when given a text item (or any text correctly describing parts of the visuals), ItemizedCLIP can often accurately attend to the correct region of interest associated with the text item, thus the attention visualization makes ItemizedCLIP naturally interpretable and makes its zero-shot decisions more trustworthy.

E.2. Region-based text retrieval

We show additional region-based text retrieval examples in Figure 17. For each brain MRI study, we manually select a sequence and a certain region vp' within the sequence (defined in local tokens of size 8x16x16), and we retrieve text items from all text items within the prospective test set (T, $\sim$100K items in total) by $\text{argmin}_{t \in T} TCSim(t, vp')$. The retrieved text items often matches well with one or more ground truth text items matched to the study, indicating ItemizedCLIP's local visual representation captures both anatomy/location information as well as pathological information.

E.3. Item differentiability

We show an additional qualitative example of IIS improving model item differentiability on brain MRI. Each brain MRI whole-study contains many 3D sequences, and sometimes a text item will refer specifically to certain 3D sequence types. In Figure 18, we show an example where two text items for a single study refers to two different types of sequences, T1 and T2. We show the top 3 sequences (out of 35 total sequences in the study) with highest total attention for each text item, on models trained with and without IIS.

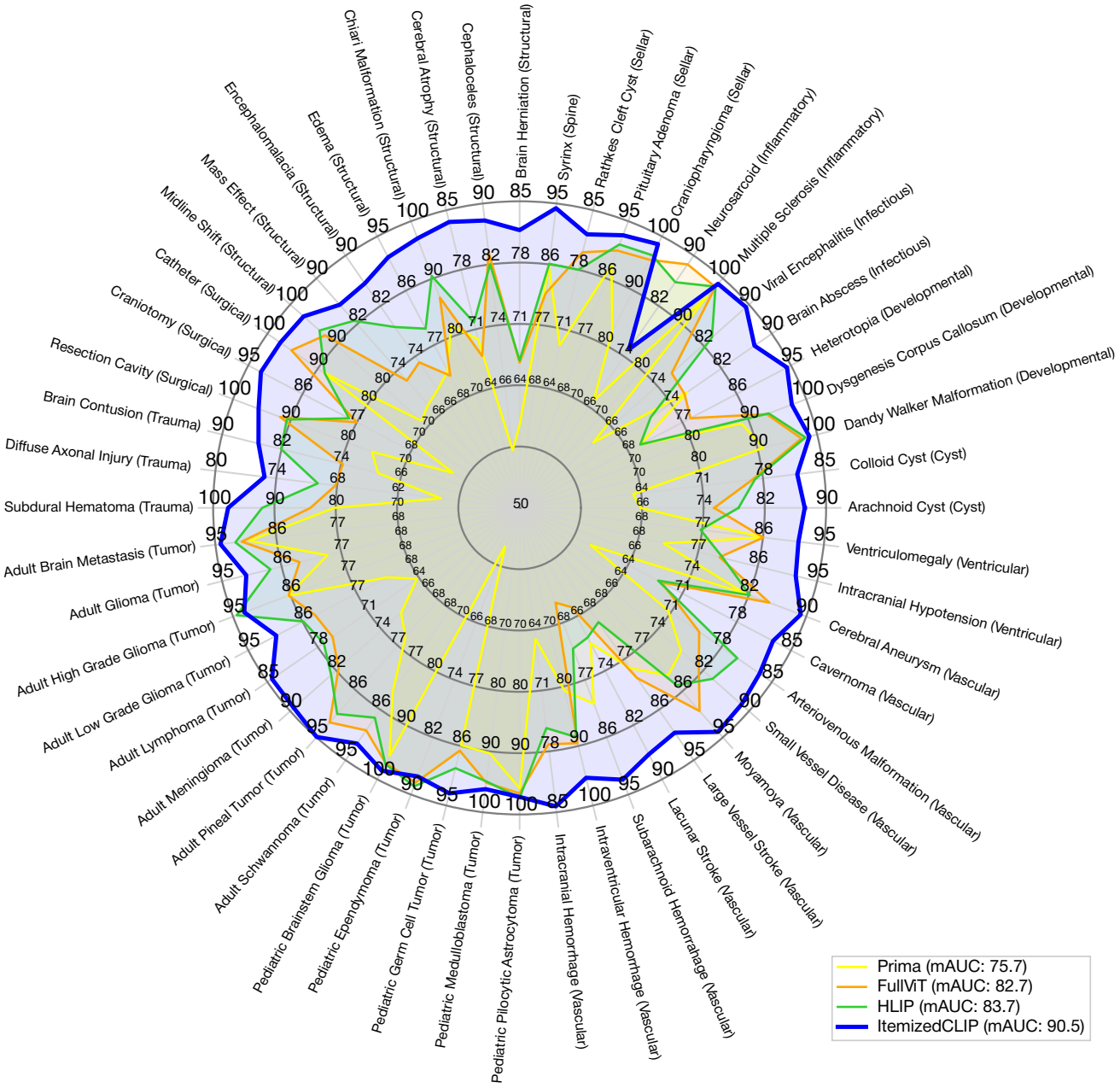


Figure 13. Radar plot on full results over all 52 tasks from Prima's prospective test set [24] (zero-shot AUC). The parenthesis after each diagnostic task is its category.

Models	Inference Type	ICH	IPH	IVH	SDH	EDH	SAH	BL-left	BL-right	MidlineShift	MassEffect	Mean AUC
FullViT [40]	Zero Shot	64.4	74.1	94.4	67.4	75.7	80.8	57.5	58.3	72.9	85.8	73.1
HLIP [40]	Zero Shot	76.5	**86.7**	**97.4**	**84.8**	88.0	87.2	72.0	75.6	75.9	86.8	83.1
ItemizedCLIP (Ours)	Zero Shot	**88.7**	86.4	**97.4**	82.1	**89.9**	**92.0**	**85.9**	**87.0**	**95.8**	**95.1**	**90.0**
FM-HeadCT [42]	Linear Probing	86.3	86.2	95.9	65.3	65.2	80.7	71.4	82.2	76.1	90.7	80.0
Google-CT [34]	Linear Probing	83.1	82.9	89.4	53.0	59.6	75.6	68.2	80.4	90.1	78.4	76.1
Merlin [2]	Linear Probing	58.9	61.6	49.8	55.0	68.9	50.3	48.3	54.9	51.4	51.4	55.1

Table 14. CQ500 per-task AUC performance.

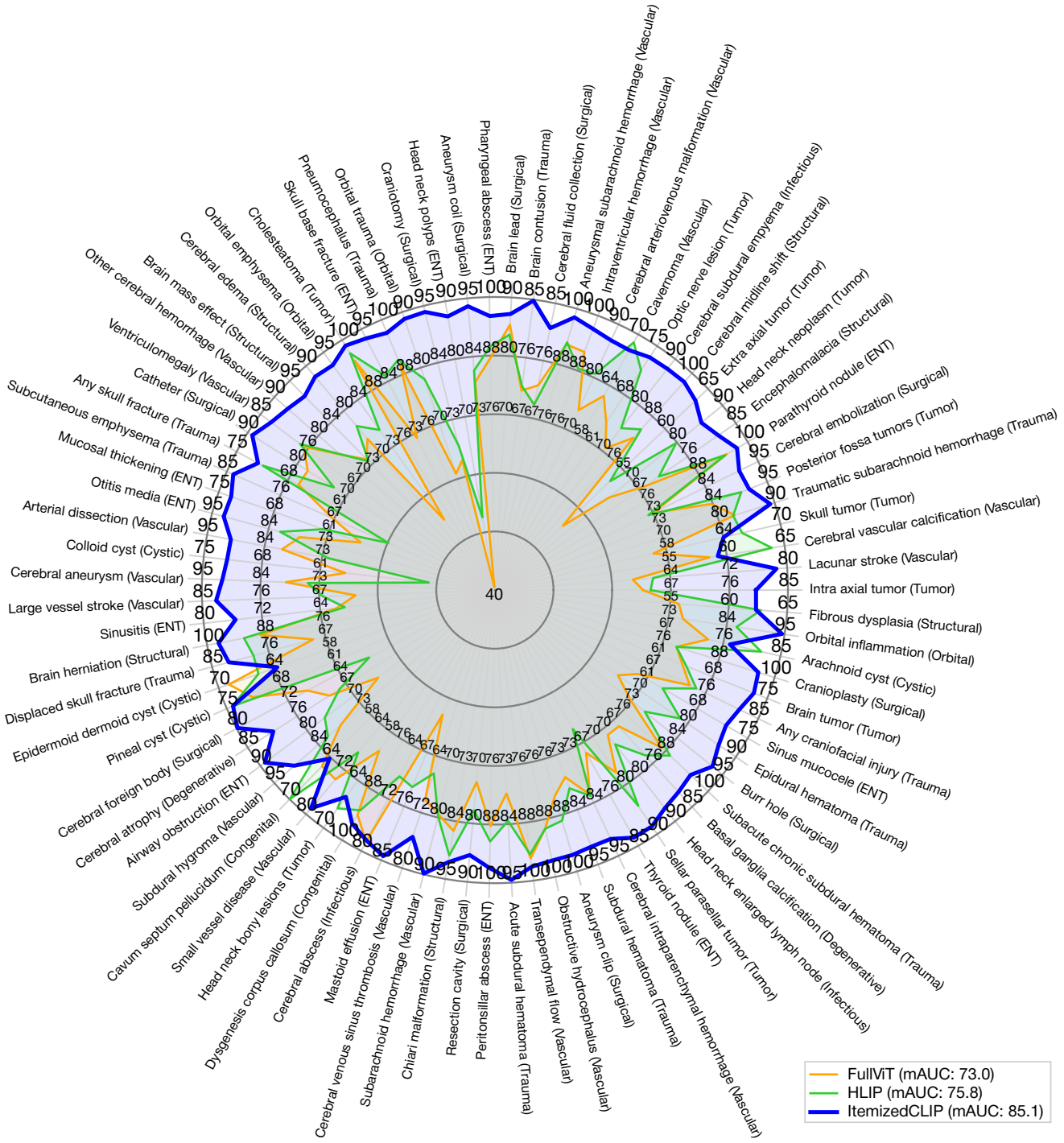


Figure 14. Radar plot on full results over all 83 tasks from HeadCT240K's prospective test set [40] (zero-shot AUC). The radar plot style is inspired by [24] and [40]. The parenthesis after each diagnostic task is its category.

We see that the model trained with IIS correctly focuses its attention on sequences with types mentioned in the text item, while the model trained without IIS always have the FLAIR and T2-SWI sequences as its top 2 focus. This example further illustrates that IIS guides the model towards better and more accurate item differentiability. (Note that FLAIR here refers to a sequence type in brain MRI, not the visual-language pre-training method FLAIR [32].)

	CT-Rate (16 tasks)									Rad-ChestCT (14 tasks)								
	HLIP-RA			HLIP-SA			ItemizedCLIP			HLIP-RA			HLIP-SA			ItemizedCLIP		
	AUC	BAcc	wF1	AUC	BAcc	wF1	AUC	BAcc	wF1	AUC	BAcc	wF1	AUC	BAcc	wF1	AUC	BAcc	wF1
Emphysema	77.0	71.3	73.8	77.3	69.2	72.0	**81.3**	**74.0**	**76.2**	76.4	71.3	72.5	74.0	67.7	69.2	**80.2**	**72.5**	**73.8**
Atelectasis	70.2	64.1	66.8	71.5	66.3	68.5	**76.8**	**68.3**	**70.6**	63.1	62.0	63.4	60.7	59.9	61.4	**67.6**	**62.8**	**64.4**
Lung nodule	59.4	57.4	57.5	61.3	59.0	58.9	**66.4**	**61.9**	**62.0**	64.3	**67.0**	**69.5**	**65.9**	63.9	67.1	**65.9**	64.5	67.6
Lung opacity	79.9	**76.6**	**76.6**	80.7	75.5	75.5	**81.1**	76.1	76.3	66.2	**61.8**	**61.8**	66.1	62.7	62.5	**66.4**	61.7	**61.8**
Pulmonary fibrotic sequela	57.8	54.6	57.0	59.1	55.9	58.1	**68.8**	**64.1**	**65.9**	84.3	78.2	81.1	**84.8**	**81.9**	**84.0**	80.7	78.9	81.5
Pleural effusion	95.8	92.9	93.3	95.5	93.0	93.3	**96.4**	**93.3**	**93.8**	**92.0**	**87.5**	**88.1**	89.3	84.1	84.9	91.5	84.7	85.6
Mosaic attenuation pattern	72.2	62.5	70.8	78.3	69.6	76.3	**86.4**	**80.9**	**84.6**	-	-	-	-	-	-	-	-	-
Peribronchial thickening	73.2	67.4	73.1	72.7	65.8	71.7	**80.5**	**74.1**	**78.4**	**64.6**	**65.9**	**73.4**	64.1	64.6	72.5	64.5	60.4	69.1
Consolidation	87.7	79.1	81.1	88.7	80.7	82.2	**90.8**	**82.9**	**84.1**	79.9	**74.8**	**78.2**	78.0	72.7	76.6	**80.4**	74.0	77.7
Bronchiectasis	72.8	63.2	70.2	72.9	70.5	75.7	**80.6**	**75.4**	**79.5**	**71.8**	**65.8**	**70.3**	70.1	63.8	68.7	69.8	62.3	67.4
Interlobular septal thickening	82.3	73.9	79.6	84.5	**75.5**	**80.7**	**85.2**	73.7	79.5	74.6	65.6	74.1	77.0	71.3	78.4	**82.4**	**78.3**	**83.3**
Cardiomegaly	89.2	80.4	83.6	91.9	81.9	84.8	**94.5**	**85.4**	**87.5**	82.1	73.2	78.0	83.5	74.9	79.4	**88.3**	**80.2**	**83.5**
Pericardial effusion	83.2	72.1	78.7	80.4	71.9	78.5	**85.9**	**80.2**	**84.4**	64.6	60.2	65.7	60.5	58.1	63.8	**66.6**	**63.9**	**68.8**
Coronary artery wall calcification	87.0	**78.8**	**79.8**	86.5	78.0	79.1	**87.1**	78.4	79.6	-	-	-	-	-	-	-	-	-
Hiatal hernia	69.7	70.1	74.3	68.8	64.3	69.7	**78.3**	**71.3**	**75.5**	60.5	59.8	67.0	60.0	56.4	64.1	**69.3**	**65.4**	**71.6**
Arterial wall calcification	85.7	77.9	78.7	89.0	81.4	81.9	**90.9**	**84.7**	**85.1**	-	-	-	-	-	-	-	-	-
Calcification	-	-	-	-	-	-	-	-	-	68.1	64.7	66.0	69.7	65.3	66.6	**72.9**	**68.0**	**69.3**
Mean	77.7	71.4	74.7	78.7	72.4	75.5	**83.2**	**76.5**	**78.9**	72.3	68.4	72.1	71.7	67.7	71.4	**74.7**	**69.8**	**73.2**

Table 15. Per-task zero-shot performance on Chest CT datasets (CT-Rate [14] test set and Rad-ChestCT [9]), comparing ItemizedCLIP to two top-performing baselines (HLIP-RA and HLIP-SA [40]).

RSICD	Mean Rank (out of 30)	Top-1 Acc	Top-5 Acc
ItemizedCLIP without Diverse Sampling	4.72	39.2	73.3
ItemizedCLIP with Diverse Sampling	**3.86**	**46.2**	**78.7**

Itemized-cc0.3M	MSCOCO				Flickr			
	I@1	I@10	T@1	T@10	I@1	I@10	T@1	T@10
ItemizedCLIP without Diverse Sampling	4.5	20.0	6.0	25.5	10.2	33.3	14.2	43.0
ItemizedCLIP with Diverse Sampling	**6.1**	**24.1**	8.1	**31.1**	**14.4**	**38.8**	**19.2**	**51.8**

Chest CT	CT-RATE (16 tasks)			Rad-ChestCT (15 tasks)		
	AUC	Balanced ACC	Weighted F1	AUC	Balanced ACC	Weighted F1
ItemizedCLIP without Diverse Sampling	**83.2**	**76.5**	**78.9**	**74.7**	**69.8**	**73.0**
ItemizedCLIP with Diverse Sampling	81.6	74.9	77.5	72.9	68.3	71.8

Table 16. Comparison between applying diverse sampling versus no diverse sampling. Diverse sampling improves performance over remote sensing and natural image tasks, but does not help in medical imaging tasks such as Chest CT.

We present more examples on item differentiability of ItemizedCLIP on Itemized-cc0.3M in Figure 19.

F. Computational resources

All experiments in this paper are conducted on a single server with 8 Nvidia L40S GPUs. Training ItemizedCLIP for whole-study brain MRI on UM220K takes 40 hours. Training ItemizedCLIP for whole-study brain CT on HeadCT240K takes 40 hours. Training ItemizedCLIP for single-sequence Chest CT takes 20 hours. Training ItemizedCLIP for Remote Sensing on RSICD takes 0.5 hours. Training ItemizedCLIP for natural images on Itemized-cc0.3M takes 6 hours. We do not observe significant computation overhead with ItemizedCLIP compared to different combinations of subsets of its components, as all the majority of computation in ItemizedCLIP as well as baselines happens within the visual and language backbones, which are kept the same for fairness of comparison.

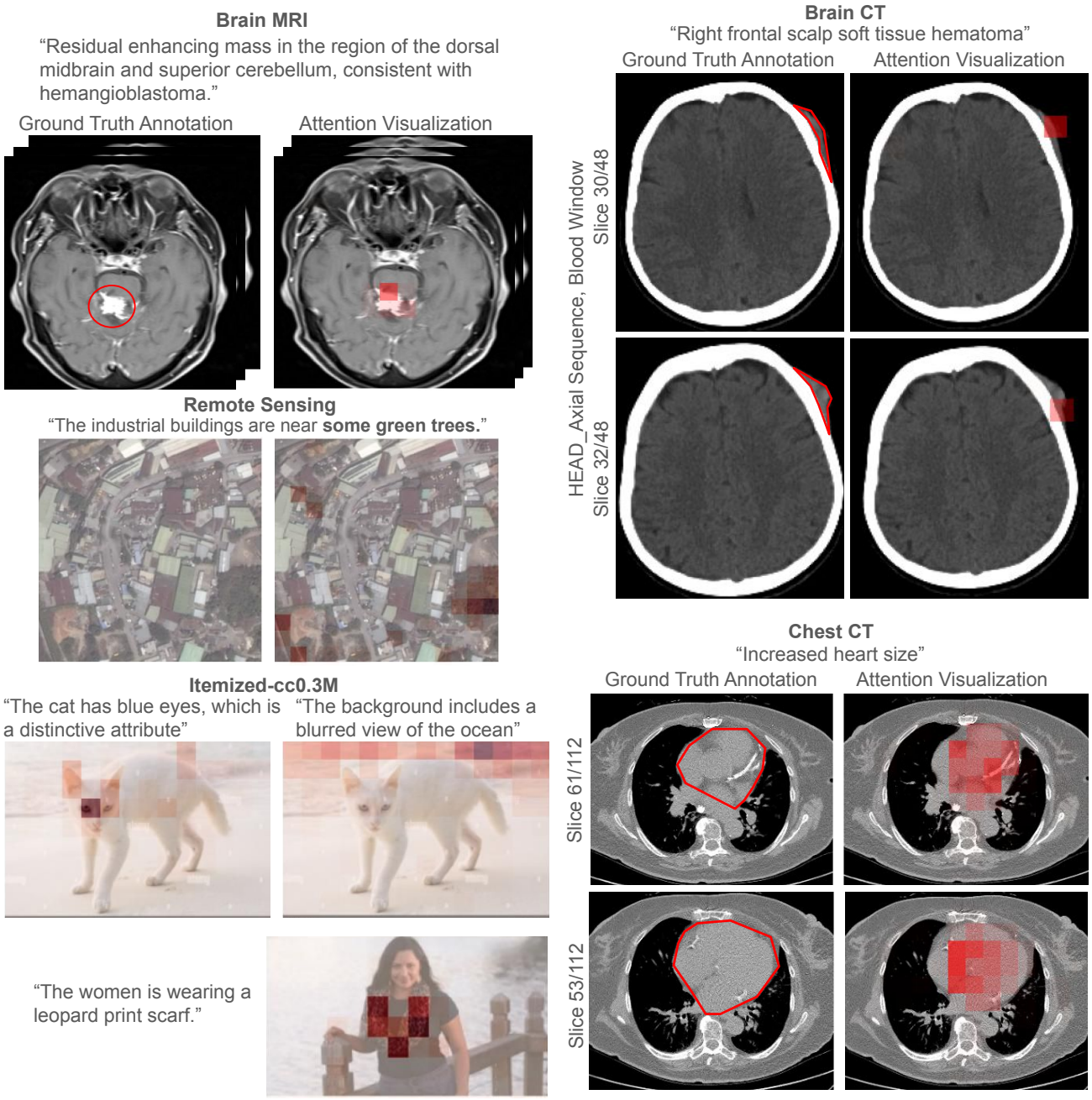


Figure 15. Additional text item attention visualization examples on all 5 evaluated domains.

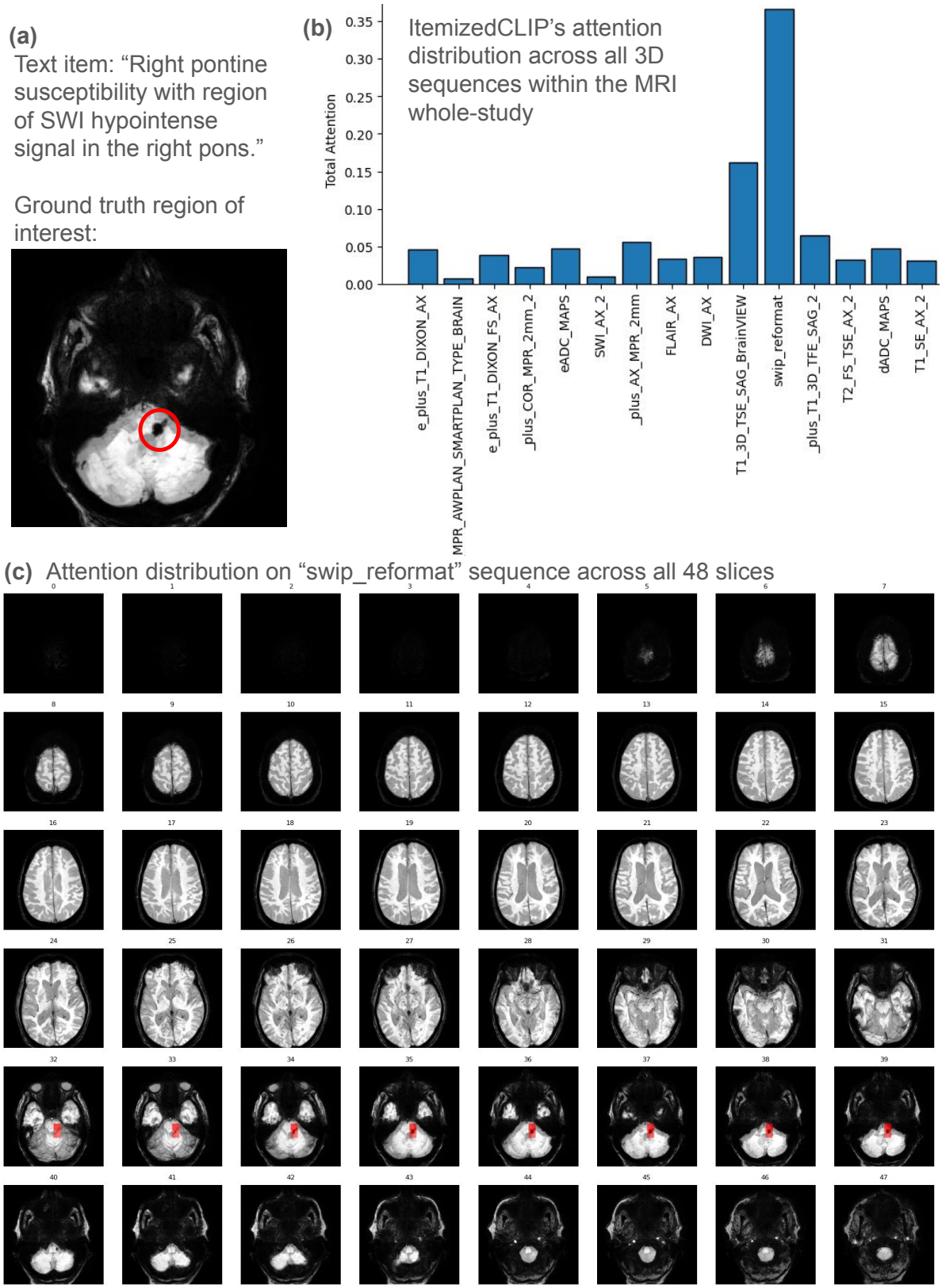
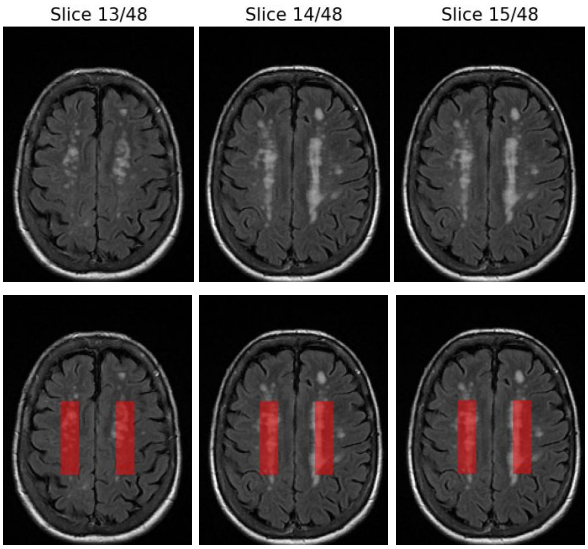
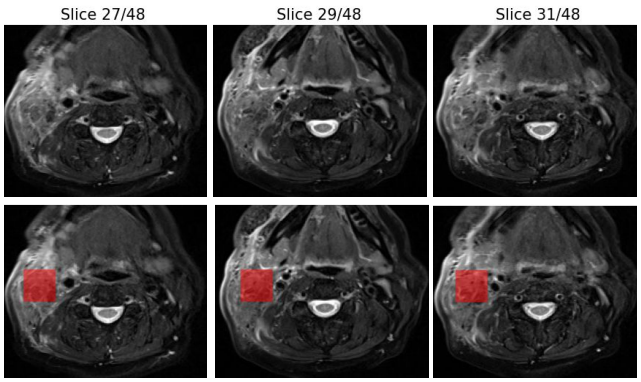


Figure 16. Visualization of attention over an entire brain MRI whole-study. **(a)** The text item and ground truth region of interest annotation about the text item in the study. **(b)** Attention distribution across all 3D sequences in the study. ItemizedCLIP correctly focused its attention on an SWI sequence (”swip_reformat”). **(c)** Attention distribution across all slices in ”swip_reformat” sequence. ItemizedCLIP correctly focused its attention over the 2 local visual tokens (each with dimensions 8x16x16) that cover the pontine susceptibility.



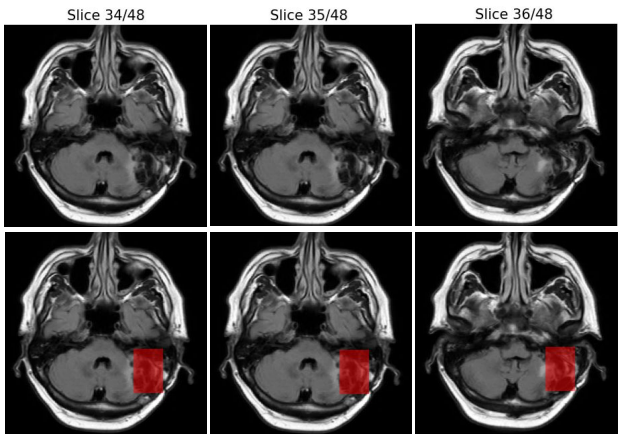
Ground truth text:
“Numerous foci of T2/FLAIR signal hyperintensity in the cerebral white matter, likely related to chronic microvascular ischemic changes”

Top Retrieved Text:
1. Multifocal confluent and patchy T2/flair hyperintensities within the periventricular and subcortical white matter.
2. Severe burden of white matter changes with patchy and confluent areas of T2 and FLAIR hyperintensity throughout the bilateral cerebral white matter.
3. Confluent periventricular and deep white matter bilateral T2 and FLAIR hyperintensities consistent with chronic demyelinating plaques.
4. New confluent patchy white matter FLAIR hyperintensities in the periventricular and subcortical white matter, corpus callosum, and bilateral middle cerebellar peduncle.



Ground truth text:
“Internal T2 hyperintensity and enhancement in the left masseter muscle, which appears thickened.”
“Extensive edema and enhancement of the anterolateral left neck superficial and deep spaces, concerning for infection.”

Top Retrieved Text:
1. Patchy enhancement of the left masseter muscle likely secondary to denervation
2. Persistent streaky enhancement and increased T2 signal within the left masseter muscle, possibly due t underlying denervation injury
3. Persistent areas of increased streaky enhancement and increased T2 signal within the left masseter muscle, possibly related to underlying denervation injury
4. New ill-defined edema and enhancement of the left buccal space



Ground truth text:
“FLAIR hyperintensity in the right posterior temporal lobe and right cerebellum, consistent with cystic encephalomalacia and gliotic changes following re-resection of the right tentorial meningioma”

Top Retrieved Text:
1. Postsurgical changes related to right mastoidectomy and suboccipital craniectomy
2. Persistent encephalomalacia in the right cerebellum
3. FLAIR signal abnormality within the right cerebellar hemisphere adjacent to the area of encephalomalacia, likely region of gliosis
4. Right cerebellar encephalomalacia and gliosis

Figure 17. Additional region-based text retrieval examples. We manually select a region in an MRI sequence (highlighted in red), then we use ItemizedCLIP to retrieve text items from all text items in prospective test set, around 100K items in total that best describe this region. We show the ground truth text item from the corresponding radiology report and top 4 retrieved items.

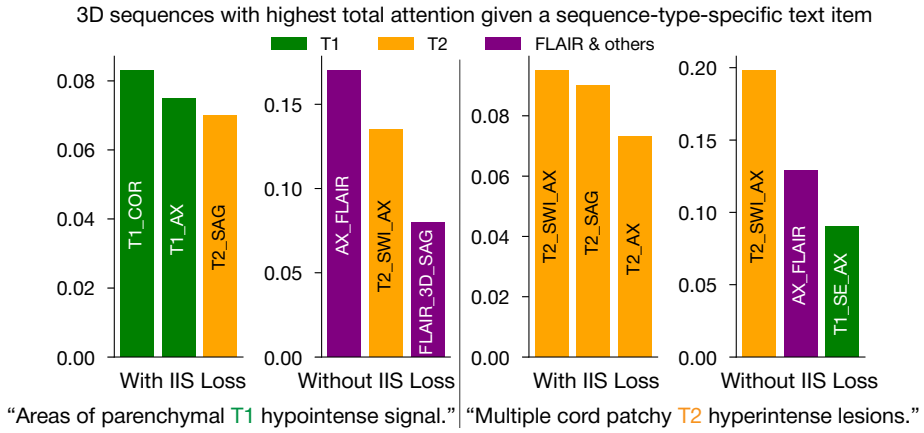


Figure 18. A brain MRI example that demonstrates item differentiability: we show the top-3 3D sequences with the highest total attention (out of 35 3D sequences in the MRI study) over 2 positive text items different text items with 2 different models, one trained with IIS and one without. The model with IIS is able to focus its top attention to the correct sequence types (T1/T2) that each text item refers to, while the model without IIS focus on the same 2 FLAIR/SWI sequences regardless of which type is mentioned in the text item. Please note that FLAIR here refers to a brain MRI sequence type rather than the visual representation learning method [32]. This example further illustrates that IIS improves item differentiability.

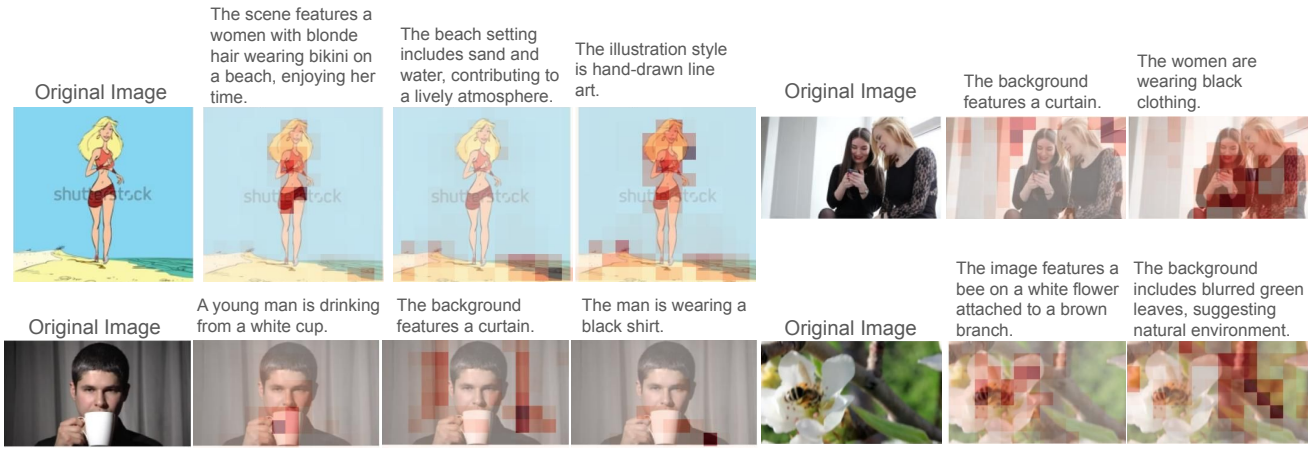


Figure 19. More item differentiability examples on Itemized-cc0.3M. In these examples, ItemizedCLIP is able to accurately visualize different text items by attending to corresponding local visual tokens.