SELF-SUPERVISED HIERARCHICAL ALIGNMENT FOR ANNOTATION-EFFICIENT CROSS-MODAL RETRIEVAL IN MEDICAL IMAGING

AI Researcher

ABSTRACT

Cross-modal retrieval between medical images and radiology reports is critical for clinical decision support but faces significant challenges due to scarce annotated paired data and the need for fine-grained semantic alignment. This research proposes a self-supervised hierarchical alignment framework that reduces annotation dependency by 70% while maintaining competitive retrieval performance. We hypothesize that modality-specific pretext tasks (image inpainting, masked language modeling) combined with hierarchical cross-modal attention will improve fine-grained mAP by 18-22% over global baselines Yang et al. (2016), particularly in zero-shot and few-shot scenarios. Our methodology integrates self-supervised contrastive learning inspired by BYOL Grill et al. (2020), hierarchical region-word attention mechanisms, and knowledge distillation from large teacher transformers to compact student architectures under 50M parameters Lan et al. (2019). Experiments will evaluate MIMIC-CXR et al. (2021) and ImageCLEFmed datasets against strong baselines including VSE++ and CLIP, measuring Recall@K, mAP, and inference efficiency (targeting 8-10x speedup) in both standard and zero-shot settings. Expected contributions include: (1) a novel annotation-efficient framework achieving >90% of supervised performance; (2) open-source pre-trained models optimized for medical deployment; and (3) comprehensive empirical guidelines for hierarchical alignment across modalities. This work advances multimodal learning in data-scarce medical domains while enabling efficient clinical deployment.

1 INTRODUCTION

Medical imaging generates an overwhelming volume of visual data that remains critically underexploited due to the semantic gap—the disconnect between pixel-level image content and high-level clinical concepts expressed in radiology reports. Cross-modal retrieval systems that enable radiologists to search image databases via natural language queries and retrieve relevant prior cases through image examples hold transformative potential for clinical decision support, educational resource discovery, and longitudinal patient monitoring Xu et al. (2023). However, developing robust medical cross-modal retrieval frameworks faces a fundamental bottleneck: the extreme scarcity of manually annotated paired data. Expert radiologist annotations are prohibitively expensive, requiring specialized training and hundreds of hours per dataset, while remaining subject to inter-observer variability and reporting inconsistencies. Large-scale benchmarks like MIMIC-CXR contain millions of chest X-ray images with corresponding free-text reports, yet creating explicit image-to-text alignment annotations for even a fraction of this data remains impractical et al. (2014). While recent vision-language transformers demonstrate promising retrieval capabilities, they rely entirely on supervised contrastive learning from fully paired annotations, limiting their scalability across diverse clinical institutions, imaging modalities (X-ray, CT, MRI, ultrasound), and evolving pathological taxonomies Xu et al. (2023).

Self-supervised learning offers a compelling pathway to circumvent annotation dependence, yet existing approaches fail to capture the rich hierarchical semantics inherent in radiological data. Methods such as adversarial cross-modal retrieval Wang et al. (2017); **?**); **?**); **?** and correspondence autoencoders Feng et al. (2014) operate primarily at global feature level, overlooking fine-grained alignments between anatomical regions and pathological descriptions. Bootstrapped self-supervised frameworks like BYOL Grill et al. (2020) learn powerful unimodal representations but lack explicit

mechanisms to align cross-modal semantics at multiple granularities. Meanwhile, hierarchical attention networks have proven effective for document classification by learning word-to-sentence and sentence-to-document relationships Yang et al. (2016), suggesting their untapped potential for aligning medical images' multi-scale visual patterns—from localized lesions to organ-level findings—with corresponding textual findings, from individual words to complete impression statements.

This research proposes a unified framework that integrates modality-specific self-supervised pretext tasks with hierarchical cross-modal alignment and knowledge distillation to achieve annotation-efficient medical cross-modal retrieval. We hypothesize that jointly learning rotation-invariant image features, masked anatomical structure prediction, inpainting tasks, and sentence ordering will reduce annotation requirements by 70% while preserving 90% of supervised baseline performance. Critically, we posit that explicit hierarchical semantic alignment—matching image regions to words, anatomical structures to phrases, and global findings to impression statements—will substantially improve fine-grained retrieval accuracy beyond global feature alignment alone. Furthermore, distilling this knowledge into lightweight architectures inspired by ALBERT Lan et al. (2019) will enable clinical deployment with sub-second retrieval latency on hospital infrastructure using models under 50M parameters.

To validate these hypotheses, we formulate four research questions: (1) How can self-supervised pretext tasks reduce dependency on manually annotated paired data while maintaining competitive retrieval performance in medical cross-modal retrieval? (2) Can hierarchical semantic alignment at multiple granularities (image regions, anatomical structures, global findings) significantly improve fine-grained cross-modal retrieval accuracy? (3) What architectural optimizations and knowledge distillation strategies enable efficient cross-modal retrieval with minimal accuracy degradation for clinical deployment? (4) How do different attention mechanisms compare in fusing multi-modal features for zero-shot and few-shot cross-modal retrieval scenarios involving unseen pathologies? Experiments will evaluate MIMIC-CXR and ImageCLEFmed using Recall@K, median rank, and mean Average Precision metrics.

This proposal is organized as follows: Section 2 reviews foundational work in self-supervised learning, cross-modal retrieval, and hierarchical attention. Section 3 details our methodological framework and hypotheses. Section 4 outlines experimental design and evaluation protocols. Section 5 discusses expected contributions and broader impacts.

2 BACKGROUND

Cross-modal retrieval between medical images and radiology reports represents a critical capability for clinical decision support, medical education, and biomedical research et al. (2021). Traditional supervised approaches rely heavily on large-scale manually annotated paired datasets, where each image is explicitly linked to its corresponding textual report Wang et al. (2017). This annotation process is prohibitively expensive in the medical domain, requiring expert radiologists to produce both imaging interpretations and fine-grained region-phrase alignments, often at a cost of $50-100 per study et al. (2014). Consequently, existing methods face severe scalability limitations, with annotation costs increasing linearly with dataset size and complexity, restricting models to single-institution datasets and limiting generalizability across diverse patient populations and imaging protocols.

Self-supervised learning has emerged as a promising paradigm to mitigate annotation dependency by leveraging unpaired data through pretext tasks. Recent advances such as Bootstrap Your Own Latent (BYOL) demonstrate that high-quality representations can be learned without negative samples by predicting transformed views of the same instance Grill et al. (2020). Similarly, ALBERT introduces cross-layer parameter sharing and sentence-order prediction to achieve efficient language pre-training with reduced annotation requirements Lan et al. (2019). These approaches have achieved remarkable success in unimodal contexts but remain underexplored for medical cross-modal alignment, where modality-specific pretext tasks (e.g., image inpainting for chest X-rays, masked anatomical structure prediction for CT scans, and disease entity ordering for reports) could unlock unprecedented data efficiency. This gap directly motivates RQ1, which investigates whether self-supervised pretext tasks can reduce annotation dependency by 70% while maintaining competitive retrieval performance.

Hierarchical semantic alignment addresses the inherent multi-scale nature of medical knowledge, where findings span from pixel-level features to global diagnostic impressions. In neuroimaging, this

hierarchy extends from individual voxels to tissue types, anatomical structures, lobes, and whole-brain networks; in histopathology, from cellular features to tissue patterns and organ-level pathology. Hierarchical Attention Networks (HAN) have demonstrated superior performance in document classification by learning word-to-sentence and sentence-to-document relationships Yang et al. (2016). While sequential variants have shown promise in recommendation systems Ying et al. (2018), their application to medical cross-modal retrieval remains nascent. Current methods predominantly rely on global feature aggregation or flat attention mechanisms, neglecting the structured anatomical and pathological hierarchies that radiologists naturally employ when interpreting studies. This limitation directly informs RQ2, which hypothesizes that hierarchical region-word attention will improve fine-grained mAP by 18-22% over global alignment baselines on MIMIC-CXR.

Existing cross-modal retrieval frameworks employ adversarial training Wang et al. (2017) or correspondence autoencoders Feng et al. (2014) to learn shared latent spaces. However, these approaches typically operate on fixed-scale representations and require full supervision. The recent proliferation of multimodal transformers Xu et al. (2023) offers new possibilities for joint representation learning, yet their computational complexity (often exceeding 500M parameters) poses deployment challenges in clinical environments with strict latency requirements (<100ms per query) and limited GPU memory (<4GB on edge devices). Knowledge distillation techniques provide a pathway to model compression, but their efficacy in preserving cross-modal retrieval performance while reducing model size by orders of magnitude remains inadequately quantified, particularly for zero-shot scenarios involving unseen pathology categories. This deployment gap motivates RQ3, which seeks architectural optimizations and distillation strategies for efficient clinical deployment.

Current attention mechanisms—including additive attention, scaled dot-product attention, and co-attention—have not been rigorously compared for medical cross-modal retrieval, particularly in few-shot regimes where query modalities may differ from training distributions. This gap directly addresses RQ4, which hypothesizes that cross-modal attention fusion will outperform simple concatenation by 12-15% in Recall@10 for zero-shot retrieval of unseen pathologies. The medical imaging community has established benchmarks like BRATS et al. (2014) for segmentation tasks and MIMIC-CXR for image-text retrieval, yet standardized evaluation protocols for annotation-efficient and hierarchical retrieval remain underdeveloped, with current metrics (Recall@K, median rank) potentially underestimating fine-grained localization accuracy.

Furthermore, while natural language processing has evolved sophisticated pre-training strategies Khurana et al. (2022), their adaptation to medical report generation and retrieval lacks systematic investigation regarding hierarchical semantic preservation and computational efficiency. This research directly addresses these gaps by proposing a unified framework that integrates modality-specific self-supervision, hierarchical semantic alignment, and efficient architecture design specifically tailored for the medical domain's unique constraints and opportunities, advancing both theoretical understanding and clinical applicability.

3 METHOD

Methodology

This research proposes a three-stage framework for annotation-efficient cross-modal medical retrieval, integrating self-supervised representation learning, hierarchical semantic alignment, and model compression. The architecture employs twin encoders for images and texts, trained initially with modality-specific pretext tasks before joint optimization for cross-modal alignment. Each stage directly addresses specific research questions from the framework.

Stage 1: Self-Supervised Unimodal Representation Learning (Addressing RQ1) To reduce annotation dependency, each modality encoder is pretrained independently using contrastive self-supervision. For medical images, we adopt a multi-task objective combining rotation prediction, image inpainting, and Momentum Contrast (MoCo)-style instance discrimination Grill et al. (2020), leveraging inherent structural priors in radiological scans. The image encoder utilizes a ResNet-50 backbone, augmented with a nonlinear projection head that maps features to a 256-dimensional latent space where differently augmented views of the same image are pulled together via InfoNCE loss (a contrastive lower bound on mutual information). For radiology reports, we implement a domain-adapted masked language modeling task inspired by ALBERT Lan et al. (2019), where 15%

of clinical terms identified through UMLS (Unified Medical Language System) Metathesaurus are masked and predicted through a 12-layer transformer encoder. This approach learns rich anatomical and pathological semantics without requiring paired data. We hypothesize that this stage will achieve >90% of supervised baseline performance using only 30% of paired annotations.

Stage 2: Hierarchical Cross-Modal Alignment (Addressing RQ2 & RQ4) Following unimodal pretraining, we introduce a hierarchical attention mechanism that aligns image regions with corresponding text phrases at three granularities: (1) fine-grained region-word alignment using a bottom-up attention module that computes similarity between anatomical regions detected by Faster R-CNN (pretrained on Open Images anatomical entity classes) and noun phrases extracted via constituency parsing; (2) structure-sentence alignment employing a cross-modal transformer that matches segmented anatomical structures to their descriptive sentences through co-attention; and (3) global-finding alignment using spatially-pooled features for overall study-level retrieval. This multi-scale design directly addresses Research Question 2. The alignment is optimized through a weighted combination of triplet losses: L_total = λL_region-word + λL_structure-sentence + λL_global-finding, where λ_i are learned parameters initialized at 0.4, 0.3, 0.3 respectively and optimized via gradient descent. Hard negative mining selects the top-5 hardest negatives per query. Drawing from hierarchical attention networks Yang et al. (2016), we incorporate a query-specific gating mechanism that dynamically reweights contributions from different semantic levels. For zero-shot scenarios (Research Question 4), we define unseen pathologies as those absent from training but present in test splits, and replace simple concatenation with a cross-modal attention fusion layer where image features attend to text embeddings through multi-head attention (8 heads), enabling better generalization.

Stage 3: Knowledge Distillation for Deployment Efficiency (Addressing RQ3) To enable clinical deployment, we distill the large teacher model (Stage 2) into a compact student architecture under 50M parameters. The student model employs a distilled transformer with 6 layers, 768 hidden dimensions, and 12 attention heads, following efficiency principles from Lan et al. (2019). We employ a layer-mapping strategy where the student's layers 1,3,5 learn to mimic the teacher's layers 3,7,11 through L2 distance and attention transfer losses. Additionally, we introduce a retrieval-specific distillation objective that preserves ranking relationships between query-response pairs through a margin ranking loss. This addresses Research Question 3, targeting 8-10x speedup with <5% accuracy degradation.

Experimental Setup We evaluate on MIMIC-CXR (377,110 chest X-ray image-report pairs) and ImageCLEFmed 2023 (12,000 medical images with captions across radiology, pathology, and dermatology modalities), following standard splits for reproducibility. Baselines include VSE++, SCAN, and CLIP fine-tuned on medical data. Evaluation metrics comprise Recall@K (K=1,5,10), median rank, and mean Average Precision (mAP). Statistical significance will be assessed via paired bootstrap testing with Bonferroni correction across five random seeds ($p<0.01$). Ablation studies will isolate contributions from each hierarchical level, pretext task, and distillation component. Implementation uses PyTorch with AdamW optimization (lr=5e-5, weight decay=0.01), trained on 4$\times$A100 GPUs for 100 epochs with early stopping on validation mAP. This integrated methodology holistically addresses annotation efficiency, fine-grained alignment, and deployment constraints for clinical cross-modal retrieval.

4 Experimental Setup

Experimental Design

Datasets and Preprocessing Experiments will utilize two publicly available medical multimodal benchmarks: **MIMIC-CXR**, comprising 377,110 chest X-ray images with associated radiology reports, and **ImageCLEFmed**, spanning five imaging modalities (X-ray, CT, MRI, ultrasound, pathology) for cross-domain evaluation. For annotation-efficiency analysis, we will construct four training subsets containing 100%, 50%, 20%, and 5% of paired annotations while preserving pathology class distributions. Chest X-rays will be resized to 224$\times$224 and normalized using modality-specific statistics. Radiology reports will be tokenized with BioWordVec embeddings and truncated/padded to 256 tokens. For hierarchical alignment, anatomical structures will be automatically segmented using a pre-trained CheXsegment model, and findings will be extracted via CheXpert labeler.

Evaluation Metrics Cross-modal retrieval performance will be assessed using **Recall@K** (K=1,5,10) for both image-to-text and text-to-image directions, **mean Average Precision (mAP)** at IoU thresholds {0.3, 0.5, 0.7}, and **median rank** of correct matches. For efficiency analysis, we will measure **inference latency** (ms/query) on NVIDIA T4 GPUs and **model size** (parameters). Statistical significance will be assessed using two-tailed paired t-tests with Bonferroni correction (α=0.05) across five random seeds.

Baseline Methods We will compare against three categories: (1) **Supervised methods**: VSE++ Wang et al. (2017), SCAN, and CLIP adapted to medical domain; (2) **Self-supervised approaches**: Image-only DINO Grill et al. (2020) and text-only ALBERT Lan et al. (2019) with late fusion; (3) **Medical-specific models**: M3AE and MedCLIP. All baselines will be re-implemented and optimized using identical training protocols (batch size 512, AdamW optimizer, learning rate 1e-4 with cosine decay) for fair comparison.

Experimental Protocol **Experiment 1 (Annotation Efficiency)**: Train the full framework with varying paired annotation percentages (100% to 5%) to validate Hypothesis 1, measuring performance degradation relative to supervised baselines. **Experiment 2 (Hierarchical Alignment)**: Conduct ablation studies removing region-level, structure-level, or global-level attention branches to isolate contributions of hierarchical semantics (Hypothesis 2). **Experiment 3 (Efficiency Optimization)**: Distill a 48M-parameter student model from a 307M-parameter teacher transformer using contrastive distillation loss with temperature τ=0.07, measuring speed-accuracy tradeoffs (Hypothesis 3). **Experiment 4 (Zero/Few-Shot)**: Evaluate retrieval on held-out pathology categories (12/50 pathologies excluded from training) comparing cross-modal attention fusion against concatenation baselines (Hypothesis 4). All experiments will employ 5-fold cross-validation, with mean$\pm$std reported. Code, model checkpoints, and detailed logs will be released for reproducibility.

5 TIMELINE

Timeline

The 24-month project is structured into six phases with explicit milestones, parallel task streams, and comprehensive risk management.

Phase 1 (Months 1-3): Foundation & Baseline Establishment Comprehensive literature review of self-supervised learning Grill et al. (2020), hierarchical attention Yang et al. (2016), and cross-modal retrieval [cite5, cite6]. Preparation and curation of MIMIC-CXR and ImageCLEFmed datasets; implementation and rigorous validation of VSE++, SCAN, and CLIP baselines. **Deliverable: Complete baseline results with Recall@K, median rank, and mAP metrics; experimental protocol finalized; target: IEEE ISBI 2025 submission.**

Phase 2 (Months 4-7): Self-Supervised Pretext Task Development (RQ1) Design and implement modality-specific pretext tasks: chest X-ray rotation prediction, image inpainting, and masked language modeling for radiology reports based on ALBERT Lan et al. (2019). Train unimodal encoders without paired annotations. **Deliverable: Pretrained image and text encoders achieving >85% of supervised performance on unimodal tasks; target: MICCAI 2025 abstract.**

Phase 3 (Months 6-10): Hierarchical Cross-Modal Alignment (RQ2) Develop multi-scale attention mechanisms aligning image regions (4×4 to 16×16 grids), anatomical structures, and global findings with corresponding text phrases. Integrate with outputs from Phase 2. **Deliverable: Hierarchical attention module achieving 18-22% mAP improvement over global alignment baselines; contribution to journal paper.**

Phase 4 (Months 9-13): Knowledge Distillation for Efficiency (RQ3) Distill large teacher transformers (>500M parameters) into compact student architectures (<50M parameters) using contrastive and relation distillation losses. **Deliverable: Student model with 8-10$\times$ inference speedup and <5% Recall@K degradation; target: AAAI 2026 submission.**

Phase 5 (Months 12-18): Comprehensive Evaluation & Ablation (RQ4) Systematic experiments across zero-shot and few-shot scenarios; component ablation studies; statistical significance testing across 5 modalities and 12 pathologies. **Deliverable: Complete performance analysis and ablation results; comprehensive empirical study for TPAMI submission.**

Phase 6 (Months 18-24): Dissemination & Open-Source Release Prepare journal/conference submissions; document clinical deployment guidelines; release pre-trained models, code, and reproducibility package. **Deliverable: Two high-impact publications, public GitHub repository, and clinical deployment whitepaper.**

Dependencies & Risk Management: Phases 2-3 run in parallel; Phase 4 depends on Phase 3 completion; Phase 5 requires outputs from Phases 3-4. **Technical Risks**: (1) Self-supervised pretext tasks may not transfer effectively to medical domain; mitigation: implement progressive fine-tuning from natural images. (2) Hierarchical alignment may suffer from noisy region proposals; mitigation: employ anatomical segmentation priors et al. (2014). (3) Knowledge distillation may cause significant performance drop; mitigation: employ intermediate representation matching and data augmentation. A 2-month contingency in Phase 6 accommodates unexpected challenges.

6 Expected Results

We anticipate four primary outcomes that directly address our research questions. First, our self-supervised framework will achieve >90% of fully supervised performance on MIMIC-CXR et al. (2021) using only 30% of paired annotations (RQ1), as modality-specific pretext tasks (image inpainting, masked language modeling) learn semantically rich unimodal representations that transfer effectively to cross-modal alignment. This would demonstrate that pretext tasks can capture anatomical and pathological semantics without costly radiologist annotations, addressing a critical bottleneck in medical AI Xu et al. (2023).

Second, hierarchical semantic alignment will yield an 18-22% improvement in fine-grained mAP over global feature baselines (VSE++, SCAN Wang et al. (2017)) (RQ2), with the greatest gains on fine-grained tasks like specific pathology localization. This confirms that multi-scale region-word attention better captures the compositional nature of radiology reports, where findings reference anatomical structures at varying granularities Yang et al. (2016).

Third, our distilled student model (<50M parameters) will achieve 8-10x inference speedup on GPU hardware with <5% degradation in Recall@K metrics (RQ3). This performance-efficiency trade-off will enable real-time retrieval in clinical PACS environments, addressing deployment constraints that have limited transformer-based methods Lan et al. (2019).

Fourth, cross-modal attention fusion will outperform concatenation baselines by 12-15% Recall@10 in zero-shot retrieval of unseen pathologies (RQ4), demonstrating robust generalization. However, we anticipate performance variance across modalities—CT-to-report retrieval may lag behind X-ray due to greater visual complexity et al. (2014).

Implications and Limitations: Success would establish annotation-efficient retrieval as a viable pathway for clinical implementation. However, alternative outcomes warrant careful interpretation: if hierarchical alignment yields diminishing returns beyond two scales, this suggests redundancy in deep semantic pyramids; if knowledge distillation suffers >10% accuracy loss, hybrid architectures rather than pure compression may be necessary; and if self-supervised pretraining fails to reduce annotation needs by 70%, this indicates domain-specific fine-tuning remains indispensable for medical vision-language tasks. Dataset-specific constraints, such as MIMIC-CXR's chest X-ray focus, may limit generalizability to other modalities like MRI et al. (2014). We will conduct rigorous ablation studies and statistical significance testing to isolate failure modes and ensure robust interpretation of each component's contribution.

7 Conclusion

This proposal introduces a self-supervised hierarchical alignment framework designed to overcome the annotation bottleneck in medical cross-modal retrieval. The research directly addresses four critical questions: reducing annotation dependency through modality-specific pretext tasks, improving fine-grained accuracy via multi-scale semantic alignment, enabling efficient deployment via knowledge distillation, and optimizing attention mechanisms for zero/few-shot scenarios. Our methodology integrates self-supervised contrastive learning Grill et al. (2020) with hierarchical region-word attention mechanisms Yang et al. (2016) within multimodal transformer architectures Xu et al. (2023), distilling knowledge into compact models under 50M parameters. We hypothesize achieving >90%

of supervised baseline performance with 70% fewer paired annotations, 18-22% mAP improvements over global alignment on MIMIC-CXR et al. (2021), and 8-10x inference speedup with <5% accuracy degradation. Broader impacts include democratizing cross-modal retrieval in resource-limited clinical settings, extending to diverse modalities like MRI and pathology imaging et al. (2014), and establishing reproducible benchmarks through open-source pre-trained models. By bridging self-supervised representation learning with fine-grained anatomical semantics, this work aims to set new standards for annotation-efficient, clinically-deployable multimodal AI systems.

REFERENCES

Bjoern Menze et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE Transactions on Medical Imaging*, 2014.

Firoj Alam et al. Findings of the association for computational linguistics: Emnlp 2021. *Unknown*, 2021.

Fangxiang Feng, Xiaojie Wang, and Ruifan Li. Cross-modal retrieval with correspondence autoencoder. *Unknown*, 2014.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Val'ko. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv (Cornell University)*, 2020.

Diksha Khurana, Aditya Koli, Kiran Khatter, and Sukhdev Singh. Natural language processing: state of the art, current trends and challenges. *Multimedia Tools and Applications*, 2022.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of languagerepresentations. *arXiv (Cornell University)*, 2019.

Shengsheng Qian, Dizhan Xue, Quan Fang, and Changsheng Xu. Adaptive label-aware graph convolutional networks for cross-modal retrieval. *IEEE Transactions on Multimedia*, 24:3520–3532, 2021a.

Shengsheng Qian, Dizhan Xue, Huaiwen Zhang, Quan Fang, and Changsheng Xu. Dual adversarial graph neural networks for multi-label cross-modal retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 2440–2448, 2021b.

Shengsheng Qian, Dizhan Xue, Quan Fang, and Changsheng Xu. Integrating multi-label contrastive learning with dual adversarial graph neural networks for cross-modal retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4794–4811, 2022.

Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalić, and Heng Tao Shen. Adversarial cross-modal retrieval. *Research Repository (Delft University of Technology)*, 2017.

Peng Xu, Xiatian Zhu, and David A. Clifton. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. *Unknown*, 2016.

Haochao Ying, Fuzhen Zhuang, Fuzheng Zhang, Yanchi Liu, Guandong Xu, Xing Xie, Hui Xiong, and Jian Wu. Sequential recommender system based on hierarchical attention networks. *Unknown*, 2018.