Image-Grounded Vision Knowledge Graphs for Topology-Aware Medical Reasoning in Abdominal CT

Anonymous CVPR submission

Paper ID *****

Abstract

Vision foundation models achieve strong performance in medical image segmentation and representation learning; however, clinical decision-making requires structured reasoning over anatomical relationships that are not explicitly encoded in voxel-level predictions or embedding spaces. In abdominal oncology, diagnosis and risk assessment depend on relational factors such as tumor burden, lesion multiplicity, vascular proximity, and cross-organ topology—characteristics that demand interpretable, constraint-aware reasoning grounded in clinical knowledge.

We introduce an image-grounded, lesion-centric Vision Knowledge Graph (VKG) framework that transforms segmentation-derived organ and tumor masks into structured relational graphs encoding anatomical containment, spatial proximity, organ topology, and quantitative tumor burden. By integrating a symbolic anatomical ontology with learned graph representations, the proposed neuro-symbolic architecture enables explicit topology-aware reasoning grounded in imaging evidence and serves as a structured reasoning layer complementary to vision and vision-language foundation models.

The VKG supports three core reasoning capabilities: (i) compositional multi-constraint retrieval across heterogeneous CT cohorts, (ii) anatomically grounded risk stratification aligned with established oncologic factors, and (iii) interpretable evidence-path generation linking predictions to structured anatomical context. To demonstrate real-world usability, we deploy an interactive web-based application that enables clinicians and researchers to perform structured queries and retrieve predictive information across three abdominal cancer CT datasets (liver, pancreas, and multi-organ cohorts), supporting explainable cohort exploration and decision support.

We evaluate topology-aware retrieval and reasoning performance against conventional baselines, including rule-based filtering, attribute-only models, multimodal embedding similarity, and relational graph neural networks. VKG reasoning consistently achieves superior structured retrieval accuracy (e.g., higher nDCG@10 under compositional clinical queries), improved predictive reasoning outcomes (higher AUROC for anatomically defined risk stratification), and stronger cross-dataset generalization under a shared ontology schema. These results indicate that explicit relational reasoning provides complementary clinical insight beyond similarity-based retrieval and embedding-driven inference.

Our findings position image-grounded knowledge graphs as a practical reasoning architecture that bridges voxel-level perception and interpretable clinical inference, enabling scalable, explainable, and interactive medical AI for cancer imaging applications.

1. Introduction

Recent advances in deep learning and vision foundation models have substantially improved tumor segmentation and representation learning in abdominal CT imaging. Self-configuring frameworks such as nnU-Net achieve strong cross-dataset robustness [7], while transformer-based architectures—including TransUNet, CoTr, and nnFormer—capture long-range contextual dependencies for accurate volumetric prediction [4, 15, 18]. As a result, modern medical vision systems can reliably detect and delineate anatomical structures and lesions across diverse clinical datasets.

However, clinical decision-making requires more than accurate perception. Diagnosis and treatment planning depend on reasoning over anatomical relationships that are not explicitly represented in voxel-level predictions or embedding spaces. In abdominal oncology, clinicians reason about tumor burden, lesion multiplicity, vascular proximity, anatomical containment, and cross-organ topological spread—relational factors central to staging systems such as BCLC and TNM. Existing pipelines typically convert segmentation outputs into tabular descriptors followed by rule-based filtering or independent predictive models, implicitly treating lesions as isolated entities and overlooking structured anatomical dependencies. Consequently, current approaches provide limited support for compositional reasoning, interpretable inference, or structured cohort exploration.

Recent multimodal representation learning methods attempt to bridge imaging and semantic understanding through embedding-based alignment. Biomedical encoders such as BioClinicalBERT, PubMedBERT, and MedCLIP enable cross-modal similarity modeling between medical images and clinical text [1, 5, 14]. While effective for similarity-based retrieval, embedding representations primarily capture statistical correlations and do not explicitly encode lesion-centric anatomical topology derived from imaging measurements. As a result, they struggle to support constraint-aware reasoning or explainable modeling of prognostic anatomical patterns.

Knowledge graphs (KGs) provide a natural representation for structured reasoning by modeling entities and relations explicitly. Prior work integrates clinical knowledge bases with imaging features for diagnosis and report generation [8, 13]; however, most medical KGs are constructed from textual ontologies rather than directly grounded in image-derived anatomical evidence. This weak coupling limits their ability to support topology-aware reasoning grounded in voxel-level observations.

In this work, we introduce an **image-grounded, lesion-centric Vision Knowledge Graph (VKG)** framework that transforms segmentation-derived organ and tumor masks into structured relational graphs encoding anatomical containment, spatial proximity, organ topology, and quantitative tumor burden. By integrating symbolic anatomical ontologies with learned graph representations, the proposed neuro-symbolic architecture establishes a structured *post-perception reasoning layer* for medical vision systems. The VKG enables explicit topology-aware reasoning grounded in imaging evidence and complements representation-driven vision and vision-language foundation models.

The proposed framework supports three reasoning capabilities: (i) compositional multi-constraint retrieval for cohort-level exploration, (ii) anatomically grounded risk stratification aligned with clinical staging factors, and (iii) interpretable evidence-path generation linking predictions to structured anatomical context. We further deploy an interactive web-based application that enables structured querying and predictive information retrieval over three abdominal cancer CT datasets (liver, pancreas, and multi-organ cohorts), supporting explainable and interactive clinical analytics.

Extensive experiments demonstrate that VKG reasoning outperforms rule-based filtering, attribute-only models, multimodal embedding retrieval, and relational graph neural networks. The proposed approach achieves improved topology-aware retrieval accuracy, stronger discrimination of anatomically defined risk patterns, and robust cross-dataset generalization under a shared lesion-centric ontology.

By bridging segmentation-derived imaging phenotypes with structured relational reasoning, this work advances medical AI from perception toward interpretable reasoning, enabling scalable and explainable cancer imaging analytics.

Contributions. The main contributions of this work are:

- **Image-grounded lesion-centric reasoning representation.** We introduce a Vision Knowledge Graph that converts segmentation-derived tumor phenotypes into structured relational entities encoding anatomical topology and quantitative tumor burden directly grounded in imaging evidence.
- **Neuro-symbolic topology-aware reasoning architecture.** We propose a reasoning framework integrating anatomical ontology constraints with learned graph representations, enabling interpretable compositional retrieval and anatomically grounded predictive inference beyond embedding-based approaches.
- **Evaluation and deployment for retrieval and reasoning.** We provide a comprehensive evaluation across three abdominal cancer CT datasets, demonstrating superior topology-aware retrieval and predictive reasoning outcomes compared with attribute-based, embedding-based, and relational GNN baselines, and deploy an interactive web application enabling real-time structured querying and predictive information retrieval.

2. Related Work

2.1. Medical Vision Models and Segmentation-Centric Pipelines

Large-scale abdominal CT benchmarks have driven significant progress in organ and tumor segmentation. The Liver Tumor Segmentation Benchmark (LiTS) established a standardized multi-center evaluation platform [2], while the FLARE challenge expanded evaluation to multi-organ abdominal CT datasets with broader anatomical coverage [10]. Architectures such as nnU-Net achieve strong cross-dataset robustness through self-configuring design [7], and transformer-based or hybrid volumetric models—including TransUNet, CoTr, and nnFormer—capture long-range contextual dependencies for improved segmentation accuracy [4, 15, 18].

Despite these advances, segmentation models primarily address perception rather than reasoning. Clinical decision-making relies on relational anatomical understanding—such as tumor burden, multiplicity, vascular proximity, and cross-organ topology—which are not explicitly represented in voxel predictions. Conventional postsegmentation pipelines therefore reduce masks to tabular descriptors, limiting compositional reasoning and structured cohort analysis.

2.2. Embedding-Based Retrieval and Multimodal Medical Representation Learning

Recent biomedical foundation models enable semantic retrieval through pretrained language and multimodal encoders. Models such as BioClinicalBERT, PubMedBERT, and SapBERT learn contextual clinical representations from large-scale corpora [1, 5, 9], while multimodal encoders such as MedCLIP align medical images and text within shared embedding spaces for cross-modal retrieval [14].

Although effective for similarity-based matching, embedding representations primarily capture statistical correlations rather than explicit anatomical structure. They do not encode lesion-centric topology derived from imaging measurements, limiting their ability to enforce structured clinical constraints or support interpretable reasoning over anatomical relationships.

2.3. Graph Representation Learning and Knowledge Graph Reasoning

Knowledge graph embedding methods such as TransE and DistMult model relational structure for entity–relation prediction [3, 16], while graph neural networks (GNNs), including GraphSAGE and relational graph convolutional networks (R-GCN), learn neighborhood-aware representations from graph topology [6, 11]. These approaches have been applied to biomedical prediction and structured diagnosis tasks.

However, most graph-based methods remain embedding-driven and optimized for predictive performance rather than explicit compositional reasoning or interpretable evidence-path analysis. Furthermore, many biomedical knowledge graphs are constructed from curated textual ontologies rather than directly grounded in image-derived anatomical measurements, resulting in weak coupling between relational reasoning and imaging evidence.

2.4. Foundation Models and Multimodal Reasoning

Transformer-based graph encoders such as Graphormer demonstrate strong capability for modeling long-range relational dependencies [17]. Concurrently, multimodal medical foundation models explore joint reasoning across imaging and language modalities. Nevertheless, existing approaches largely emphasize representation alignment and embedding learning, with limited focus on constructing structured, image-grounded relational representations that explicitly encode anatomical topology and disease extent.

2.5. Knowledge Graphs for Clinical Decision Support

Knowledge graphs have increasingly been used to integrate structured biomedical knowledge with imaging data for diagnosis and report generation [8, 13]. Prior systems typically rely on external knowledge bases or curated ontologies, rather than deriving relational structure directly from imaging measurements. Consequently, they provide limited support for topology-aware reasoning grounded in segmentation-derived anatomical evidence.

2.6. Positioning of Our Contribution

In contrast to segmentation-centric pipelines [4, 7, 15, 18], embedding-based retrieval approaches [1, 14], and representation-driven KG or GNN models [3, 6, 11], we propose an image-grounded, lesion-centric Vision Knowledge Graph (VKG) framework that directly converts segmentation-derived tumor measurements into typed relational entities encoding anatomical containment, spatial proximity, organ topology, and quantitative tumor burden.

Our neuro-symbolic framework integrates anatomical constraints with learned graph representations to enable: (i) topology-aware compositional retrieval, (ii) anatomically grounded predictive reasoning aligned with oncologic staging factors, and (iii) cross-dataset reasoning under a shared ontology schema. By explicitly grounding relational inference in imaging-derived evidence, VKG establishes a structured post-perception reasoning layer that bridges voxel-level analysis and interpretable medical reasoning in abdominal CT.

3. Method

3.1. Overview

Our goal is to move from perception to structured reasoning in medical vision systems. Given abdominal CT volumes, we construct an image-grounded, lesion-centric *Vision Knowledge Graph* (VKG) that supports topology-aware retrieval, interpretable predictive reasoning, and cross-dataset generalization.

As illustrated in Figure **??**, segmentation masks are first obtained using AUSAM [12]. From these masks, we extract geometric and relational descriptors encoding lesion–organ topology. These entities and relations are assembled into a structured VKG that integrates symbolic anatomical constraints with learned graph representations. The resulting architecture forms a neuro-symbolic reasoning layer operating directly on segmentation-derived imaging evidence.

3.2. Image-Grounded Anatomical Feature Extraction

Organ and tumor masks are obtained using AUSAM [12]. For each 3D CT scan, let $L = \{l_1, \ldots, l_n\}$ denote tumor instances and $O = \{o_1, \dots, o_m\}$ denote segmented organs.

We derive lesion-centric anatomical descriptors:

- Lesion volume $V(l_i)$
- Host organ via volumetric containment
- Coverage ratio $C(l_i) = \frac{V(l_i)}{V(o_j)}$
- Lesion multiplicity
- Minimum surface-to-surface distance to adjacent organs or vessels (vascular proximity)
- Inter-lesion spatial proximity

These measurements encode the structural extent and form of disease and form the foundation for explicit relational modeling.

3.3. Vision Knowledge Graph Construction

We construct a lesion-centric VKG $G = (\mathcal{V}, \mathcal{E})$, where nodes represent imaging entities and edges encode topology derived directly from segmentation geometry.

- **Node types:** Scan, Organ, Lesion.
- **Relation types:** Containment (lesion → organ), Proximity (lesion ↔ organ or vessel), Inter-lesion adjacency, Organ topology, Quantitative attribute relations.

This representation transforms voxel-level perception outputs into a structured, relational graph that supports compositional reasoning.

3.4. Neuro-Symbolic Reasoning Architecture

The VKG integrates complementary symbolic and neural components:

- **Symbolic layer:** a typed anatomical ontology enforcing relational constraints (containment, adjacency, proximity).
- **Neural layer:** graph neural network (GNN) embeddings learned from VKG topology.

Hybrid inference combines constraint satisfaction with learned relational representations, enabling interpretable and topology-aware reasoning grounded in anatomical structure.

3.5. Topology-Aware Retrieval

Given a structured query q defined by anatomical constraints, retrieval evaluates relational predicates over the VKG.

Let $s(v, q)$ denote symbolic constraint satisfaction and $\phi(v)$ denote neural similarity. We compute ranking as:

$$R(v, q) = \alpha\, s(v, q) + (1 - \alpha)\, \phi(v),$$

where α balances explicit topology reasoning and embedding similarity.

This neuro-symbolic ranking enables multi-constraint retrieval beyond attribute filtering or embedding-only matching.

3.6. Predictive Anatomical Reasoning

For scan-level predictive reasoning, we operate on subgraphs G_s:

$$h_s = \text{READOUT}\,(\text{GNN}(G_s))\,.$$

The representation h_s encodes topology-aware anatomical context. A prediction head estimates Anatomical Risk Stratification (ARS) categories derived from relational features such as tumor burden, multiplicity, vascular proximity, and cross-organ spread.

Relational modeling captures contextual dependencies unavailable to attribute-only predictors.

3.7. Cross-Dataset Reasoning via Shared Ontology

All datasets are aligned to a unified lesion-centric ontology with consistent node semantics and normalized attributes. This enables models trained on one cohort to reason over another without dataset-specific feature engineering, supporting topology-aware transfer across heterogeneous CT datasets.

3.8. Training and Inference

Graph encoders are trained using supervised ARS objectives with optional relational link-prediction losses to preserve structural consistency.

At inference time, a single VKG supports both structured retrieval and predictive reasoning, yielding a unified reasoning framework for interactive querying and risk assessment.

3.9. Computational Complexity

Let N_s denote scans, N_o organs, and N_l lesions. Feature extraction scales linearly with voxel count. Graph construction requires $\mathcal{O}(N_s(N_o + N_l))$ for containment edges and $\mathcal{O}(N_s N_l N_o)$ for proximity edges.

Because anatomical graphs are sparse, VKGs remain memory-efficient and support near real-time retrieval and inference.

3.10. Evaluation Overview

We evaluate whether lesion-centric Vision Knowledge Graph (VKG) reasoning improves clinically meaningful *topology-aware retrieval* and *anatomically grounded risk stratification* beyond attribute-only, embedding-based, and graph-representation baselines. Our goal is to isolate the value of *explicit relational reasoning* rather than gains attributable solely to higher-capacity feature representations.

Controlled setting. Segmentation masks from benchmark abdominal CT datasets are treated as structured imaging evidence encoding lesion geometry and anatomical context. To reduce confounding from upstream variability, masks are refined using AUSAM, enabling consistent extraction of containment, proximity, and topology relations.

Therefore, our evaluation focuses on the reasoning layer and downstream inference, rather than segmentation accuracy.

Capability-tier evaluation. We adopt a *capability-tier* framework in which progressively richer representations are introduced while keeping training and evaluation protocols fixed:

- *T1: Attribute-only* — tabular lesion phenotypes (volume/coverage/multiplicity/proximity),
- *T2: Semantic* — PubMedBERT text embeddings,
- *T3: Multimodal* — BiomedCLIP visual embeddings and zero-shot scores,
- *T4: Graph* — relational embeddings learned using GraphSAGE,
- *T5: VKG (Ours)* — topology-aware reasoning features derived from the VKG (evidence-path completeness, topology diversity, adjacency density).

This cumulative design enables controlled estimation of the marginal contribution of relational structure and reasoning signals. In addition to aggregate metrics, we perform **query-level analyses** using representative multi-constraint queries to characterize *when* each tier is sufficient. These analyses reveal three distinct retrieval regimes: (i) *metric-driven* queries (e.g., proximity-only) where T1 suffices, (ii) *perceptual* queries (e.g., coverage patterns) where T3 provides large gains, and (iii) *structural reasoning* queries involving containment/topology where T5 is required to recover all relevant cases.

All experiments use 5-fold stratified cross-validation with strict train/test separation to prevent patient-level leakage.

Evaluation axes. We assess four complementary capabilities: (1) structured retrieval under compositional constraints, (2) anatomical risk stratification (ARS) prediction, (3) cross-dataset transfer without retraining, and (4) mechanistic ablations to quantify which VKG components drive performance.

3.11. Datasets

Experiments use three publicly available abdominal CT cohorts spanning increasing anatomical complexity and topology richness:

- *LiTS:* 118 contrast-enhanced CT scans with liver and liver tumor annotations [2]. This single-organ cohort provides a controlled setting for lesion-centric reasoning.
- *Pancreas Tumor CT:* 281 CT scans with pancreas and tumor annotations. Compared to LiTS, this cohort exhibits higher anatomical variability and boundary ambiguity due to pancreas morphology.
- *FLARE 2023:* 422 multi-organ abdominal CT scans with annotated abdominal organs [10], providing rich cross-organ topology for evaluating multi-hop reasoning.

These cohorts enable within-domain evaluation and realistic distribution shifts, while maintaining a unified lesion-centric ontology for consistent representation across datasets.

3.12. Unified VKG Representation

All cohorts are instantiated under a shared lesion-centric ontology that represents each CT study as a relational graph. The VKG contains four primary node types (*Scan*, *Organ*, *Tumor*, *Image*) and relations encoding anatomical containment, spatial proximity, visualization linkage, and inter-organ topology.

Topology-augmented edges are derived directly from segmentation geometry (e.g., organ adjacency and tumor–organ proximity), enabling multi-hop traversal across anatomical structures and reasoning over clinically meaningful context (e.g., lesion location relative to neighboring organs) rather than local appearance alone.

Each scan is represented using the cumulative capability-tier hierarchy (T1–T5), allowing controlled comparison under identical training conditions. Table 1 summarizes the tier composition, and Fig. **??** illustrates progressive VKG construction.

3.13. Structured Retrieval Evaluation

Systematic query benchmark. To evaluate topology-aware retrieval under clinically meaningful constraints, we construct a reproducible structured query benchmark grounded in segmentation-derived phenotypes. For each cohort, we generate *200 fixed multi-predicate queries* using a controlled sampling protocol with a fixed random seed, yielding:

$$3 \text{ datasets} \times 200 \text{ queries} = 600 \text{ total queries}. \quad (451)$$

Each query is a conjunction of 1–3 constraints over lesion-centric phenotypes derived from AUSAM-refined masks: (i) tumor burden (max lesion volume, coverage ratio), (ii) lesion multiplicity, (iii) organ containment, and (iv) proximity to adjacent organs or vessels.

Query modes. To control reasoning difficulty, queries are stratified into five clinically motivated modes:

- *Coverage (60/200)*: tumor burden–focused queries,
- *Severity (50/200)*: multiplicity and proximity–oriented queries,
- *Risk (40/200)*: high/borderline anatomical risk strata,
- *Perplexity (30/200)*: rare phenotype configurations based on a rarity score,
- *Hard (20/200)*: multi-constraint compositional topology queries.

Continuous thresholds are sampled from dataset-specific percentile ranges (25th–75th) to avoid extreme outliers

Table 1. Capability-tier feature hierarchy used in the unified VKG representation. Higher tiers cumulatively include all lower-tier features.

Tier	Representation	Feature Sources	Dim
T1	Attribute-only	Tumor volume, coverage, lesion statistics	7
T2	Semantic	PubMedBERT text embeddings	39
T3	Multimodal	BiomedCLIP visual embeddings + zero-shot scores	79
T4	Graph	GraphSAGE relational embeddings	111
T5	VKG (Ours)	Topology-aware reasoning features	131

Table 2. Structured retrieval performance (mean nDCG@10 over 200 queries per dataset).

Model	LiTS	Pancreas	FLARE
SQL (Rule-based)	0.360	0.369	0.349
T1 Attribute-only	0.683	0.507	0.508
T2 Semantic	0.691	0.479	0.514
T3 Multimodal (CLIP)	0.802	0.897	0.568
T4 GraphSAGE	0.798	0.905	0.566
T5 VKG (Ours)	**0.857**	**0.937**	**0.797**

while maintaining balanced relevance sets. All benchmark queries are released to ensure full reproducibility.

Evaluation protocol. Retrieval is evaluated with 5-fold cross-validation:

$$200 \text{ queries} \times 5 \text{ folds} \times 3 \text{ datasets} \times 5 \text{ tiers} = 15{,}000$$

retrieval evaluations. A scan is labeled relevant if and only if it satisfies all query constraints. Ranking quality is measured using *nDCG@10* (primary), with Precision@10 and Recall@10 as secondary metrics. Results are reported as mean ± standard deviation across queries and stratified by query mode (Fig. **??**).

Representative query analysis. Beyond aggregate performance, we report representative query cases to illustrate *capability emergence*. We observe that single-constraint metric queries (e.g., proximity-only) are already solved by T1, while coverage-driven queries exhibit large jumps at T3, indicating that organ-level coverage patterns are better captured by visual representations than by tabular summaries. Most importantly, three-constraint queries that include containment/topology exhibit a plateau from T1–T4 and are only resolved when VKG reasoning features are introduced at T5, demonstrating the necessity of explicit anatomical topology and evidence-path features for compositional retrieval.

Baselines. We compare against representative retrieval paradigms: *SQL rule-based filtering*, *T1* attribute models, *T2* text embedding similarity, *T3* CLIP-based representations, *T4* GraphSAGE embeddings, and *T5* VKG reasoning.

3.13.1. Retrieval Results

Table 2 reports mean nDCG@10 across the full 200-query benchmark per dataset. SQL filtering achieves modest performance (0.349–0.369), reflecting limitations of purely logical matching. Learned models improve retrieval with progressive gains across tiers.

VKG reasoning (T5) achieves the highest performance across all cohorts, with the largest gain on FLARE where multi-organ topology is richest. Improvements are most pronounced in *Hard* and *Perplexity* modes, indicating that explicit relational reasoning is especially beneficial under compositional and rare-case retrieval scenarios. Query-level analyses further confirm that (i) metric-only queries show negligible benefit beyond T1, (ii) coverage-focused queries can be solved at T3 due to perceptual representations, and (iii) topology/containment-driven multi-constraint queries require T5 to recover all relevant scans.

3.14. Anatomical Prognostic Risk Modeling

We evaluate whether richer representations improve identification of anatomically defined prognostic risk patterns beyond tabular tumor descriptors. Rather than requiring longitudinal survival labels, we define risk strata from established survival-associated anatomical factors reported in abdominal oncology, including tumor burden, multiplicity, vascular proximity, and cross-organ spread.

Anatomical Risk Stratification (ARS). Using VKG-derived measurements, we construct an ARS score reflecting structural disease extent: (i) total tumor volume and coverage ratio, (ii) lesion multiplicity, (iii) minimum distance to major vessels (vascular invasion proxy), and (iv) cross-organ adjacency/topological spread. Patients are categorized into low/intermediate/high risk groups using percentile-based thresholds computed from these VKG features.

Evaluation protocol. ARS group prediction is evaluated across tiers (T1–T5) using AUROC under 5-fold stratified

Table 3. Average ARS prediction performance (AUROC) across capability tiers.

Model	LiTS	Pancreas	FLARE
T1 Attribute-only	0.663	0.571	0.662
T2 Semantic	0.639	0.544	0.634
T3 Multimodal	0.749	0.733	0.788
T4 GraphSAGE	0.741	0.742	0.793
T5 VKG (Ours)	**0.762**	**0.757**	**0.859**

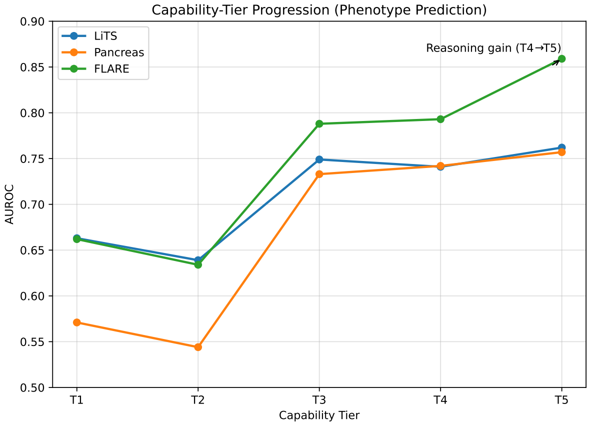

Figure 1. Capability-tier progression for Anatomical Risk Stratification (ARS).

cross-validation. Average AUROC improves monotonically:

- LiTS: 0.663 → 0.762
- Pancreas: 0.571 → 0.757
- FLARE: 0.662 → 0.859

Gains are largest on FLARE, where multi-organ topology contributes substantially to risk characterization. Figure 1 shows consistent progression from T1 to T5, indicating that explicit relational reasoning provides complementary prognostic signal beyond embeddings alone.

Overall, results indicate that image-grounded relational modeling enhances identification of anatomically defined prognostic risk patterns without requiring longitudinal outcome labels.

3.15. Cross-Dataset Transfer

To evaluate robustness under distribution shift, models trained on one cohort are evaluated on another without retraining, using the shared lesion-centric ontology schema. This setting tests whether representations capture dataset-invariant anatomical structure rather than dataset-specific appearance statistics.

Table 4 reports transfer performance (nDCG@10) for structured retrieval. VKG reasoning achieves the strongest transfer:

Table 4. Cross-dataset transfer performance (nDCG@10). Training on source, evaluating on target without retraining.

Train → Test	**T1**	**T4**	**T5 (Ours)**
LiTS → FLARE	0.299	0.304	**0.595**
FLARE → LiTS	0.373	0.527	**0.814**
LiTS → Pancreas	0.527	0.638	**0.910**

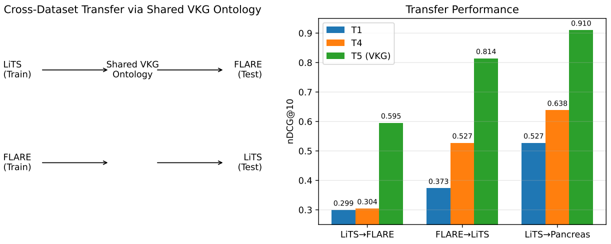

Figure 2. Cross-dataset transfer via a shared VKG ontology (left) and transfer performance (right).

Table 5. Ablation analysis on FLARE.

Configuration	nDCG@10	AUROC
Full VKG	**0.936**	**0.853**
–Proximity	0.903	0.851
–Topology	0.651	0.792
–Ontology	0.940	0.855
–Reasoning	0.658	0.785

- LiTS→FLARE: 0.595
- FLARE→LiTS: 0.814
- LiTS→Pancreas: 0.910

These results suggest that ontology-aligned VKG reasoning encodes structurally meaningful anatomical relationships that generalize beyond dataset-specific imaging distributions.

3.16. Ablation Analysis

Ablations on FLARE isolate contributions of proximity, topology, ontology, and reasoning components (Table 5). Removing explicit reasoning produces the largest degradation, confirming that performance gains arise from relational inference rather than embeddings alone.

3.17. Evidence-Path Validation

We evaluate interpretability by validating *evidence paths* produced for retrieval and ARS prediction. Each output is accompanied by ranked relational paths connecting *Tumor* nodes to *Organ* and *Scan* context through typed relations (containment, adjacency, proximity, topology).

Explanation quality is measured by: (i) **Clinical Validity**, the fraction of paths anatomically plausible under ontol-

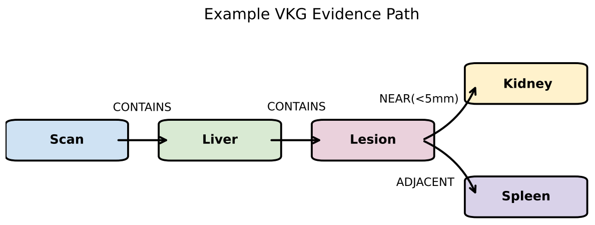

Figure 3. Example VKG evidence path connecting lesions to anatomical context through typed relations.

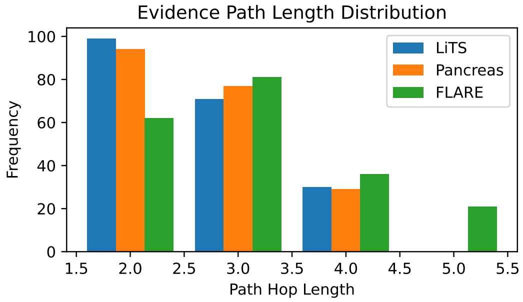

Figure 4. Evidence-path hop length distribution. FLARE exhibits more multi-hop reasoning due to richer topology.

ogy constraints; and (ii) **Constraint Satisfaction**, the fraction satisfying all query/prediction constraints. Table 6 reports high validity (0.816–0.925), indicating that reasoning outputs are grounded in anatomically consistent relations rather than similarity-only matching.

3.18. Efficiency Analysis

We measure preprocessing cost, indexing overhead, and inference latency on FLARE using a single GPU and standard CPU resources. VKG construction and indexing are performed *offline*, while prediction runs *online*, reflecting realistic deployment where graphs are constructed once and reused.

Table **??** summarizes runtime. VKG construction adds negligible overhead (5.05 ms/scan), indexing is linear in dataset size, and online inference requires 2.87 ms/scan with end-to-end latency 30.14 ms/scan, supporting real-time cohort exploration.

Stage	**Total [s]**	**ms/scan**
VKG Construction (Offline)	2.13	5.05
Graph Indexing (Offline)	9.37	22.21
Prediction (Online)	1.21	2.87
Pipeline Latency	–	**30.14**

Table 6. Evidence-path validation across datasets.

Dataset	**Clinical Validity**	**Constraint Sat.**	**Avg. Hops**	**% Multi-hop (≥3)**
LiTS	0.900	0.870	2.6	41%
Pancreas	0.816–0.880	0.850	2.8	46%
FLARE	0.925	0.890	3.1	58%

4. Three Representative Case Studies

To systematically evaluate the reasoning capabilities of the proposed Vision Knowledge Graph (VKG) framework, we present three representative case studies spanning increasing levels of anatomical and clinical complexity. Each study examines how progressively richer representations support both *topology-aware retrieval* and *clinically meaningful phenotype prediction*.

The case studies highlight complementary aspects of the proposed post-perception reasoning architecture:

1. **Compositional Constraint Satisfaction.** We analyze scenarios in which only neuro-symbolic VKG reasoning can jointly verify multiple anatomical conditions, demonstrating reasoning capabilities unavailable to feature- or embedding-based models.
2. **Progressive Emergence of Multimodal Reasoning.** We show how reasoning capability develops incrementally as representations evolve from tabular attributes to multimodal embeddings, relational graph encoding, and ultimately symbolic constraint verification.
3. **Clinically Realistic Ranking Correction.** We illustrate how structured multimodal reasoning reorganizes retrieval rankings to prioritize anatomically consistent cases, improving alignment with clinically relevant disease phenotypes and treatment-related risk assessment.

Collectively, these case studies provide: (i) theoretical evidence that neural similarity alone is insufficient for compositional anatomical reasoning, (ii) empirical validation of the staged multimodal architecture, and (iii) practical clinical motivation by demonstrating improved retrieval and phenotype inference in realistic oncology scenarios.

Together, the results position VKG as a structured reasoning layer that bridges perception-driven representations and clinically interpretable decision support.

4.1. Case Study 1: When Only VKG Reasoning Solves the Task

Dataset and Clinical Query. Using the LiTS liver tumor dataset, we analyze a clinically motivated retrieval scenario reflecting advanced hepatocellular carcinoma (HCC) risk patterns. The query targets patients exhibiting a high-risk anatomical configuration characterized by:

- **Large tumor burden**, defined as high cumulative lesion volume relative to liver size;
- **Close proximity to major vascular structures** (e.g., portal vein or hepatic veins), indicating elevated risk of vascular invasion;
- **Complex multi-lesion topology**, including spatially dispersed or bilobar tumor distribution.

This configuration reflects clinically significant disease progression, often associated with advanced staging (e.g., BCLC stage B/C) and poorer prognosis. Identifying such cases requires simultaneous reasoning over tumor size, anatomical containment within liver segments, vascular adjacency, and multi-lesion spatial organization.

Unlike single-attribute filtering, this query requires *compositional anatomical reasoning* across interdependent constraints.

Retrieval Performance Across Reasoning Tiers. Performance progresses as T1 (3/6), T2 (2/6), T3 (3/6), T4 (3/6), and T5 (6/6).

All tiers plateau until VKG reasoning (T5), which achieves perfect retrieval of clinically consistent high-risk configurations.

Why Simpler Models Fail. Tabular features (T1) capture scalar measurements such as total tumor volume or minimum vessel distance, but treat these attributes independently. As a result, cases satisfying only partial criteria are incorrectly retrieved.

Text and vision embeddings (T2–T3) improve semantic similarity but cannot enforce logical conjunction of structured clinical predicates. Graph neural networks (T4) encode relational connectivity yet lack explicit constraint verification over vascular adjacency and bilobar spread.

VKG as Clinical Constraint Verifier. Only VKG reasoning evaluates structured anatomical predicates jointly:

$$(\text{tumor burden}) \wedge (\text{vascular proximity}) \wedge (\text{bilobar topology}),$$

through ontology-grounded constraint satisfaction:

$$R(v,q) = \alpha\, s(v,q) + (1-\alpha)\, \phi(v),$$

where $s(v,q)$ verifies symbolic anatomical conditions and $\phi(v)$ captures neural similarity.

Extension to Phenotype Prediction. Beyond retrieval, we evaluate whether the same VKG representations improve prediction of clinically defined high-risk phenotypes (e.g., advanced-stage vs. early-stage disease). Using scan-level VKG embeddings as input to a classifier, we observe consistent improvements in discrimination of advanced-stage patterns compared to tabular and embedding-only baselines.

Importantly, cases correctly retrieved by VKG reasoning correspond to anatomically coherent high-risk configurations, which also exhibit improved phenotype prediction performance. This indicates that structured compositional reasoning enhances both retrieval accuracy and downstream clinical risk stratification.

Clinical Significance. In liver oncology, vascular invasion and bilobar disease distribution strongly influence treatment planning, including resection eligibility, transarterial chemoembolization (TACE), or systemic therapy. Failure to jointly evaluate these factors leads to clinically misleading retrieval results and inaccurate risk characterization.

Key Insight. Neural similarity alone is insufficient for clinically meaningful anatomical reasoning or phenotype prediction. Structured knowledge-graph reasoning enables explicit verification of multi-factor oncologic risk patterns and improves both retrieval and risk stratification performance.

4.2. Case Study 2: Progressive Emergence of Multimodal Reasoning

Dataset and Clinical Query. Using the pancreas CT dataset, we evaluate a clinically motivated retrieval and phenotype assessment task reflecting anatomical risk evaluation in pancreatic cancer. The query targets cases exhibiting a high-risk tumor configuration defined by:

- **High tumor coverage within the pancreas**, indicating substantial organ involvement;
- **Correct anatomical containment**, ensuring lesions originate within pancreatic tissue rather than adjacent organs;
- **Close proximity to critical surrounding structures** (e.g., superior mesenteric vessels or duodenum), associated with surgical complexity and resectability risk.

These factors directly influence tumor staging, surgical eligibility, and prognosis. The task therefore requires *joint reasoning over geometric measurements, anatomical relationships, and spatial topology*, making it suitable for evaluating progressive reasoning capabilities.

Performance Progression Across Reasoning Tiers.

$$T1 \rightarrow T2 \rightarrow T3 \rightarrow T4 \rightarrow T5$$
$$: \quad 2 \rightarrow 1 \rightarrow 8 \rightarrow 8 \rightarrow 9.$$

The progression reveals a gradual emergence of reasoning capability as increasingly structured representations are introduced.

Interpretation of Each Tier.

- **T1 (Tabular Features).** Scalar measurements such as tumor volume and centroid distance capture basic statistics but ignore anatomical topology and organ context, resulting in low retrieval precision (2/10) and weak phenotype discrimination.
- **T2 (Text Embeddings).** Radiology-text similarity introduces semantic alignment but lacks geometric grounding. Semantic correlations introduce noise without structural guarantees, decreasing performance (1/10).
- **T3 (CLIP Visual Embeddings).** Vision-language representations encode global spatial configuration, implicitly capturing organ shape, tumor placement, and contextual anatomy. This produces a large structural improvement (8/10), indicating that visual embeddings recover latent anatomical organization essential for both retrieval and phenotype recognition.
- **T4 (Relational GNN Encoding).** Graph neural networks propagate relational context among organs and lesions, stabilizing rankings through neighborhood reasoning. However, embeddings remain continuous approximations and cannot explicitly verify logical constraints, causing performance to plateau (8/10).
- **T5 (VKG Neuro-Symbolic Reasoning).** Vision Knowledge Graph reasoning introduces ontology-grounded relations and explicit constraint evaluation, jointly verifying containment, coverage, and proximity predicates. This resolves the remaining ambiguous case and achieves the final improvement (9/10).

Extension to Phenotype Prediction. We further evaluate whether progressively enriched representations improve prediction of clinically meaningful phenotypes such as high-risk versus resectable tumor configurations. Phenotype classification performance follows a similar progression: early tiers provide weak discrimination, while multimodal embeddings substantially improve separability, and VKG reasoning yields the most consistent predictions by explicitly validating anatomical constraints underlying clinical risk definitions.

Notably, cases correctly retrieved at higher tiers correspond to anatomically coherent disease patterns, demonstrating alignment between retrieval quality and phenotype prediction accuracy.

Scientific Insight. Reasoning capability emerges progressively rather than abruptly. Each modality contributes a distinct representational advance:

- tabular features provide scalar filtering,
- text introduces semantic alignment,
- visual embeddings recover implicit spatial structure,
- relational GNNs encode contextual dependencies,
- VKG reasoning enables explicit compositional inference.

Thus, multimodal perception alone is insufficient; structured reasoning is required to formalize anatomical constraints for reliable inference.

Architectural Validation. The observed progression empirically validates the proposed layered architecture:

$$\text{Perception} \rightarrow \text{Multimodal Embedding} \rightarrow \text{Relational Encoding} \rightarrow \text{Symbolic Reasoning}. \tag{1}$$

Each tier increases representational structure, culminating in ontology-grounded constraint satisfaction supporting both retrieval and predictive inference.

Clinical Significance. In pancreatic oncology, vessel proximity and organ containment strongly influence surgical eligibility and prognosis. A system incapable of reasoning over these relationships may retrieve anatomically irrelevant cases or produce unstable risk predictions. The progressive gains observed here demonstrate how multimodal reasoning aligns both retrieval outcomes and phenotype prediction with clinically meaningful anatomical structure.

4.3. Case Study 3: Clinically Realistic Ranking Correction

Dataset and Clinical Query. Using the pancreas CT dataset, we evaluate a clinically realistic retrieval scenario reflecting surgical risk assessment in pancreatic cancer. The query targets cases exhibiting:

- **High tumor coverage within the pancreas**, indicating substantial organ involvement;
- **Close proximity to major vascular structures** (e.g., superior mesenteric artery/vein), suggesting elevated surgical complexity and potential vascular encasement.

In clinical practice, both tumor burden and vascular proximity critically influence resectability decisions, margin status prediction, and treatment planning. Thus, the retrieval task must prioritize anatomically valid high-risk configurations rather than scans that merely satisfy scalar thresholds.

Performance Improvement. Retrieval performance improves from

$$T1 = 4/14 \quad \longrightarrow \quad T3+ = 10/14.$$

This substantial gain reflects a fundamental reorganization of ranking behavior.

Ranking Behavior Under Tabular Features (T1). With scalar attribute filtering alone, scans with large tumors are ranked highly even when the lesions are spatially distant from critical vessels. Because tabular features treat coverage and proximity independently, cases satisfying only one criterion may receive inflated scores. As a result, anatomically irrelevant scans dominate the top retrieval set, despite appearing numerically similar.

Multimodal Structural Correction (T3 and Beyond). Introducing visual multimodal embeddings reshapes the ranking. CLIP representations encode global spatial configuration, including tumor orientation, organ boundaries, and vessel adjacency patterns. Consequently:

- scans violating vascular proximity constraints are demoted,
- anatomically consistent pancreas–vessel configurations rise in ranking,
- clinically meaningful structural patterns dominate the top results.

This correction does not merely increase aggregate precision; it reorganizes the retrieval ordering to align with true anatomical structure.

Clinical Significance. In pancreatic oncology, proximity to major vessels determines borderline resectability, neoadjuvant therapy eligibility, and operative strategy. A retrieval system that ranks anatomically inconsistent cases highly would mislead cohort exploration and bias downstream analysis.

The multimodal reasoning tiers correct this misalignment by encoding spatial context unavailable in tabular representations.

Key Insight. This case demonstrates that multimodal structural reasoning improves clinically realistic ranking behavior by integrating global spatial organization into similarity scoring.

Unlike Case Study 1 (which isolates strict compositional constraint satisfaction), this scenario highlights practical clinical utility: structured multimodal reasoning prevents anatomically misleading retrieval results in real-world oncology settings.

4.4. Summary of Insights

The three case studies collectively provide complementary evidence for the role of structured reasoning in medical vision-language systems, highlighting how neuro-symbolic inference extends beyond perception-driven representations.

- **Compositional Constraint Satisfaction.** Vision Knowledge Graph (VKG) reasoning enables joint verification of multiple anatomical conditions, supporting structured clinical queries that cannot be solved by feature-based filtering or embedding similarity alone.
- **Progressive Emergence of Reasoning.** Reasoning capability increases systematically as representations evolve from scalar attributes to multimodal embeddings, relational graph encoding, and ultimately ontology-grounded symbolic constraint evaluation, demonstrating a staged transition from perception to reasoning.

- **Clinically Meaningful Ranking Correction.** Structured reasoning reorganizes retrieval rankings to prioritize anatomically consistent cases, improving alignment between retrieved cohorts and clinically relevant disease phenotypes and risk characteristics.

Collectively, these findings validate the proposed framework as a *structured post-perception reasoning layer* that bridges visual perception and clinical inference by integrating multimodal representations with explicit knowledge-graph constraint satisfaction for both medical image retrieval and phenotype prediction.

5. Interactive Neuro-Symbolic Reasoning for Retrieval and Prediction

To demonstrate the practical impact of the proposed Vision Knowledge Graph (VKG) reasoning framework, we integrate quantitative evaluation with an interactive Streamlit-based application supporting structured clinical queries and phenotype prediction.

5.1. Integrated Quantitative Summary

Table 7 summarizes retrieval performance across the three representative case studies. Each task requires increasing levels of anatomical reasoning complexity.

Across all scenarios, performance improvements occur only when structured neuro-symbolic reasoning is introduced. These results demonstrate that embedding similarity alone is insufficient for compositional anatomical queries.

5.2. Interactive Query and Prediction Interface

We developed a Streamlit-based application that enables users to perform:

- Topology-aware clinical retrieval via structured queries,
- Phenotype prediction for high-risk anatomical configurations,
- 3D visualization of organ and lesion geometry,
- Inspection of image-grounded VKG structures and evidence paths.

The interactive system directly reflects the quantitative results in Table 7. For each case study, users can visualize the anatomical configuration in 3D, inspect the corresponding VKG, formulate structured constraints, and observe ranked retrieval results or predicted phenotypes alongside evidence-path explanations.

5.3. Unified Interpretation

Together, the quantitative evaluation and interactive system demonstrate that the proposed framework functions as a structured post-perception reasoning layer. Multimodal embeddings provide geometric and relational representations, while VKG reasoning enables explicit compositional constraint satisfaction and clinically interpretable inference.

This unified design supports both topology-aware retrieval and phenotype prediction, bridging visual perception and clinical decision support.

Table 7. Integrated evaluation across case studies. Retrieval performance (Relevant@K) comparing baseline (T1) and VKG reasoning (T5).

Case Study	**Clinical Task**	**Baseline**	**VKG (T5)**
CS1 (LiTS)	High-risk liver phenotype retrieval	3/6	**6/6**
CS2 (Pancreas)	Coverage + proximity reasoning	2/10	**9/10**
CS3 (Pancreas)	Clinical ranking correction	4/14	**10/14**

Figure 5. **Interactive VKG reasoning system.** (a) 3D visualization of CT-derived organ and lesion geometry. (b) Image-grounded Vision Knowledge Graph (VKG) with typed anatomical relations (containment, proximity, topology). (c) Structured clinical query interface supporting compositional constraint specification. (d) Retrieval and phenotype prediction output with interpretable evidence-path explanations.

5.4. Key Findings

- *Capability-tier progression:* Performance improves consistently from T1→T5, indicating that structured relational reasoning contributes complementary anatomical signal beyond attributes and embeddings.
- *Reasoning drives gains:* Ablations show the largest degradation when explicit reasoning is removed, confirming that improvements are not explained by embeddings alone.
- *Topology matters most in complex anatomy:* Gains are largest on FLARE, where multi-organ topology increases the need for multi-hop reasoning.
- *Robustness under domain shift:* VKG reasoning generalizes strongly across datasets, supporting the hypothesis that ontology-aligned representations capture dataset-invariant anatomical structure.
- *Practical efficiency:* The pipeline achieves near real-time latency (∼30 ms/scan), enabling interactive cohort exploration and decision support.

References

[1] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, 2019. 2, 3

[2] Patrick Bilic, Patrick Christ, Hongwei Bran Li, Eugene Vorontsov, Avi Ben-Cohen, Georgios Kaissis, Adi Szeskin, Colin Jacobs, Gabriel Efrain Humpire Mamani, Gabriel Chartrand, et al. The liver tumor segmentation benchmark (lits). *Medical image analysis*, 84:102680, 2023. 2, 5

[3] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2013. 3

[4] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021. 1, 2, 3

[5] Yu Gu, Richard Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23, 2021. 2, 3

[6] William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 3

[7] Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021. 1, 2, 3

[8] Yingshu Li, Zhanyu Wang, Yunyi Liu, Lei Wang, Lingqiao Liu, and Luping Zhou. Kargen: Knowledge-enhanced automated radiology report generation using large language models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 382–392. Springer, 2024. 2, 3

[9] Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. Self-alignment pretraining for biomedical entity representations. In *Proceedings of NAACL*, 2021. 3

[10] Jun Ma, Yao Zhang, Song Gu, Cheng Ge, Ershuai Wang, Qin Zhou, Ziyan Huang, Pengju Lyu, Jian He, and Bo Wang. Automatic organ and pan-cancer segmentation in abdomen ct: the flare 2023 challenge. *arXiv preprint arXiv:2408.12534*, 2024. 2, 5

[11] Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *European Semantic Web Conference (ESWC)*, 2018. 3

[12] Suraj Sood, Saeed Alqarni, Syed Jawad Hussain Shah, and Yugyung Lee. Ausam: Adaptive unified segmentation anything model for multi-modality tumor segmentation and enhanced detection in medical imaging. *Knowledge-Based Systems*, 319:113588, 2025. 3

[13] Song Wang, Mingquan Lin, Tirthankar Ghosal, Ying Ding, and Yifan Peng. Knowledge graph applications in medical imaging analysis: a scoping review. *Health data science*, 2022:9841548, 2022. 2, 3

[14] Zifeng Wang, Zhen Wu, Dheeraj Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. In *EMNLP*, 2022. 2, 3

[15] Yutong Xie, Jianpeng Zhang, Chunhua Shen, and Yong Xia. Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 171–180. Springer, 2021. 1, 2, 3

[16] Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases. In *International Conference on Learning Representations (ICLR)*, 2015. 3

[17] Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. Do transformers really perform badly for graph representation? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 3

[18] Hong-Yu Zhou, Jiansen Guo, Yinghao Zhang, Xiaoguang Han, Lequan Yu, Liansheng Wang, and Yizhou Yu. nnformer: Volumetric medical image segmentation via a 3d

transformer. *IEEE transactions on image processing*, 32: 4036–4045, 2023. 1, 2, 3