# TAVR-VLM: Risk-Conditioned Causal Grounding for Hallucination-Resistant Report Generation

Zhixiang Lu[1,4], Xiwei Liu[2], Sifan Song[1], Changkai Ji[3], Anh Nguyen[4], Jionglong Su[1], Imran Razzak[2], and Jinfeng Wang[1,5]

[1] Xi'an Jiaotong-Liverpool University
[2] Mohamed bin Zayed University of Artificial Intelligence
[3] Shanghai Jiao Tong University
[4] University of Liverpool
[5] Kunming University of Science and Technology
zhixiang@liverpool.ac.uk

**Abstract.** Transcatheter Aortic Valve Replacement (TAVR) planning requires meticulous multimodal reasoning. However, adapting Multimodal Large Language Models (MLLMs) to this high-stakes domain is severely impeded by diagnostic hallucinations, where generated text lacks anatomical grounding. To address this, TAVR-VLM is introduced: a novel framework featuring Risk-Conditioned Causal Grounding Attention (R-CGA) that instantiates a model-internal "Risk → Region → Word" structural grounding pathway. R-CGA compresses multimodal inputs into a causal risk bottleneck, purifying dense visual features into a global risk mask. During autoregressive generation, a support-projected causal consistency objective constrains token-level grounding within the risk-defined support mask. Evaluated on M$^3$TAVR, a comprehensive 1,482-patient cohort, TAVR-VLM establishes a new state-of-the-art. It achieves an AUROC of 0.896, boosts CIDEr to 0.936, and drastically reduces the hallucination rate to 8.1%, thereby improving interpretability for evidence-based surgical AI.

**Keywords:** Vision-Language Models · Causal Grounding · Hallucination Mitigation · Transcatheter Aortic Valve Replacement.

## 1 Introduction

Transcatheter Aortic Valve Replacement (TAVR) has established itself as the standard-of-care for severe aortic stenosis across the full spectrum of surgical risk [7,16]. Despite its clinical success, procedural outcomes remain highly sensitive to the precision of preoperative planning. Subtle inaccuracies in the assessment of 3D CT annular geometry or 2D echocardiographic hemodynamics can precipitate catastrophic complications, including annular rupture, coronary obstruction, or significant paravalvular leak [18,9]. To ensure safety and reproducibility, the Valve Academic Research Consortium (VARC-3) emphasizes the necessity of evidence-grounded evaluation [18]. Consequently, an ideal AI-driven

TAVR system must transcend simple risk scoring to provide structured clinical reports and actionable recommendations explicitly grounded in multimodal imaging evidence.

The rapid evolution of multimodal large language models (MLLMs) has introduced the potential for generalist medical assistants capable of joint image-text interpretation [8]. While radiology report generation has progressed significantly using Transformer-based architectures on large-scale datasets [6,15], adapting these models to TAVR-grade clinical reasoning reveals a critical technical bottleneck. TAVR planning requires a rigorous coupling between global procedural risk states and localized anatomical findings. Contemporary medical MLLMs frequently exhibit diagnostic hallucinations, where the model generates plausible but non-existent clinical findings or recommendations unsupported by the underlying imagery [5,13]. In the context of structural heart interventions, such hallucinations are unacceptable as they directly compromise surgical decision-making. We posit that hallucinations persist in medical MLLMs due to the absence of explicit causal constraints [14,22]. Traditional models typically generate text by attending to dense visual tokens without ensuring that each reported clinical entity is anchored to risk-relevant evidence [13,15]. In professional TAVR workflows, clinicians utilize a top-down reasoning paradigm: they first establish a global risk hypothesis, then systematically interrogate specific anatomical regions to seek supporting evidence, and finally synthesize these insights into a structured plan [12,13]. This observation motivates our core design principle: Risk should serve as a causal bottleneck that governs the visual perception and linguistic output of the generator.

To operationalize this principle, we introduce **Risk-Conditioned Causal Grounding Attention (R-CGA)**, a top-down closed-loop module that integrates risk stratification with multi-granularity grounding for TAVR planning. R-CGA first predicts a global risk distribution and compresses it into a learnable, causal bottleneck. This bottleneck is then utilized to purify dense visual features into a global risk mask, ensuring the model encourages the model to focus on risk-relevant evidence relevant to the predicted risk state. Finally, we introduce a causal consistency supervision mechanism with a stop-gradient regularizer. This ensures that the token-level grounding maps for all generated clinical entities are spatially contained within the global risk-evidence region, thereby transforming risk prediction into a structural prior that mitigates hallucinations. Our main contributions are summarized as follows:

- We formalize TAVR assessment as a risk-conditioned structured generation task, identifying hallucination as a failure of the top-down causal link between global risk states and local anatomical evidence.
- We propose a three-stage architecture comprising a causal risk bottleneck, risk-to-region evidence selection, and region-to-word token grounding to enforce a strict "Risk $\rightarrow$ Region $\rightarrow$ Word" reasoning hierarchy.
- We conduct comprehensive experiments on the $M^3$TAVR, including risk prediction, report generation, hallucination evaluation, spatial grounding, ablation analysis and qualitative grounding analysis.

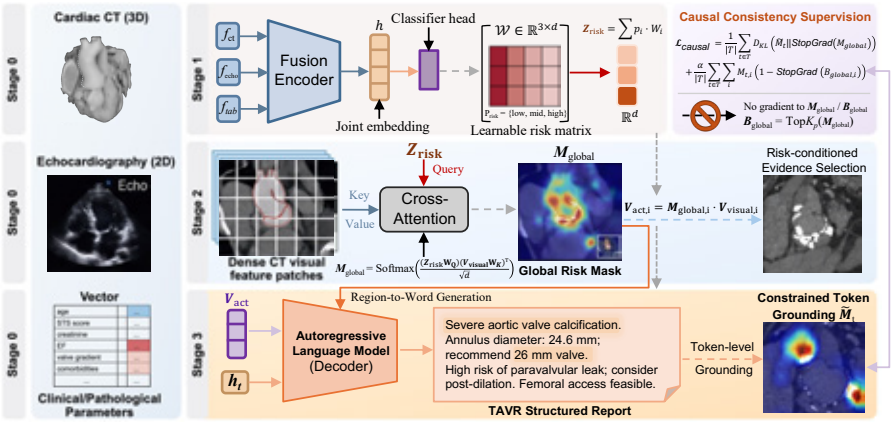



**Fig. 1. Overview of the TAVR-VLM architecture.** R-CGA implements a model-internal "Risk → Region → Word" structural grounding pathway. **Stage 1** encodes 3D CT, echocardiography clips, and clinical parameters into a joint embedding, predicts a risk distribution $P_{\text{risk}}$, and constructs a causal risk bottleneck $Z_{\text{risk}} = P_{\text{risk}}^{\top} W$. **Stage 2** uses $Z_{\text{risk}}$ as the query and CT visual tokens as key/value features to obtain a global risk mask $M_{\text{global}}$, from which a risk-defined support mask $B_{\text{global}} = \text{TopK}\rho(M\text{global})$ is derived. **Stage 3** generates the structured report from risk-activated visual features $V_{\text{act}}$ while projecting raw token attention $M_t$ into $B_{\text{global}}$ to obtain the constrained token mask $\widetilde{M}_t$. The causal consistency loss aligns $\widetilde{M}_t$ with the global risk mask and penalizes attention outside the risk-defined support.

## 2 Methodology

### 2.1 Problem Formulation and Causal Framework

We study TAVR pre-operative decision support as a risk-conditioned multimodal generation problem. Given multimodal inputs $\mathcal{X} = \{X_{\text{CT}}, X_{\text{Echo}}, X_{\text{tab}}\}$ (3D CT, 2D echocardiography, and clinical tables), our objective is to learn a joint distribution $P(Y, r \mid \mathcal{X})$. Here, $r \in \{l, m, h\}$ denotes the discrete procedural risk level, and $Y = (y_1, \ldots, y_T)$ represents the autoregressive structured report. To mitigate diagnostic hallucinations, we depart from standard multi-task learning by introducing a strict top-down causal constraint: Risk → Region → Word. We operationalize this via the **Risk-Conditioned Causal Grounding Attention (R-CGA)** module, which enforces that every clinically salient entity in $Y$ is intrinsically grounded in risk-relevant anatomical evidence (Fig. 1).

### 2.2 Risk-Conditioned Causal Grounding Attention (R-CGA)

R-CGA implements the causal hierarchy through three interconnected stages, acting as an information bottleneck that filters irrelevant visual noise prior to text generation.

*Stage 1: Causal Risk Bottleneck.* Modality-specific encoders map inputs into a unified latent space, extracting dense visual tokens $V \in \mathbb{R}^{N\times d}$. A lightweight fusion network computes a joint patient embedding to predict the risk distribution $P_{\text{risk}} = [p_l, p_m, p_h]^\top$. Instead of passing raw embeddings to the generator, we introduce a learnable risk matrix $W \in \mathbb{R}^{3\times d}$. The **Causal Risk Bottleneck** is computed as the expected prototype under $P_{\text{risk}}$:

$$Z_{\text{risk}} = P_{\text{risk}}^\top W \in \mathbb{R}^d. \tag{1}$$

$Z_{\text{risk}}$ serves as a distribution-aware macro "commander" state, encapsulating the minimal sufficient clinical evidence required to justify the predicted risk profile.

*Stage 2: Risk-to-Region Evidence Selection.* Let $V_{\text{visual}} \in \mathbb{R}^{N\times d}$ denote the dense CT visual tokens used for spatial grounding. The risk bottleneck $Z_{\text{risk}}$ acts as the query, while $V_{\text{visual}}$ serves as key/value features. The global risk mask is obtained by cross-attention [4,11]:

$$M_{\text{global}} = \text{Softmax}\left(\frac{(Z_{\text{risk}}W_Q)(V_{\text{visual}}W_K)^\top}{\sqrt{d_k}}\right) \in \Delta^{N-1}. \tag{2}$$

The risk-activated visual features are then computed by token-wise gating:

$$V_{\text{act},i} = M_{\text{global},i} \cdot V_{\text{visual},i}, \qquad i = 1, \ldots, N. \tag{3}$$

This separates the soft global risk mask used for visual activation from the support mask used for strict token containment.

*Stage 3: Region-to-Word Token Grounding.* During autoregressive generation, the decoder hidden state $h_t$ computes a raw token-level attention map over the original CT visual tokens:

$$M_t = \text{Softmax}\left(\frac{(h_t W_Q^{(t)})(V_{\text{visual}}W_K^{(t)})^\top}{\sqrt{d_k}}\right) \in \Delta^{N-1}. \tag{4}$$

Let $\mathcal{T}$ denote the set of token indices corresponding to clinically salient entities, such as "calcification", "LVOT". For each $t \in \mathcal{T}$, $M_t$ represents the raw visual evidence used by the decoder before causal support projection.

### 2.3 Support-Projected Causal Consistency and Optimization

The global risk mask $M_{\text{global}}$ identifies the anatomical region that is relevant to the predicted risk state. To enforce that clinically salient tokens are grounded within this region, we derive a risk-defined support mask:

$$B_{\text{global}} = \text{TopK}_\rho(M_{\text{global}}), \tag{5}$$

where $\mathrm{TopK}_\rho(\cdot)$ retains the top $\rho$ proportion of visual tokens and returns a binary support mask. We then project the raw token grounding map $M_t$ into this risk-defined support:

$$\widetilde{M}_t = \frac{M_t \odot \mathrm{StopGrad}(B_{\mathrm{global}})}{\sum_i M_{t,i}\mathrm{StopGrad}(B_{\mathrm{global},i}) + \epsilon}. \tag{6}$$

By construction,

$$\mathrm{supp}(\widetilde{M}_t) \subseteq \mathrm{supp}(B_{\mathrm{global}}), \quad \forall t \in \mathcal{T}. \tag{7}$$

This support projection provides the formal containment property that the raw KL alignment alone cannot guarantee. The final **Support-Projected Causal Consistency Loss** $\mathcal{L}_{\mathrm{causal}}$ contains two terms:

$$\begin{aligned}\mathcal{L}_{\mathrm{causal}} = &\frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}} \mathbb{D}_{\mathrm{KL}}\big(\widetilde{M}_t \parallel \mathrm{StopGrad}(M_{\mathrm{global}})\big) \\ &+ \frac{\alpha}{|\mathcal{T}|}\sum_{t\in\mathcal{T}}\sum_i M_{t,i}\big(1 - \mathrm{StopGrad}(B_{\mathrm{global},i})\big).\end{aligned} \tag{8}$$

The first term aligns the constrained token grounding distribution with the global risk mask. The second term penalizes raw token attention that leaks outside the risk-defined support. The StopGrad operator is applied only to the causal consistency branch, so that token grounding is optimized to obey the risk-defined region without allowing the token loss to expand or diffuse the global risk support. The end-to-end multi-task objective is jointly optimized:

$$\mathcal{L} = \mathcal{L}_{\mathrm{risk}} + \mathcal{L}_{\mathrm{LM}} + \lambda_{\mathrm{rec}}\mathcal{L}_{\mathrm{rec}} + \lambda_{\mathrm{causal}}\mathcal{L}_{\mathrm{causal}}, \tag{9}$$

where $\mathcal{L}_{\mathrm{risk}}$ is the cross-entropy for risk prediction, $\mathcal{L}_{\mathrm{LM}}$ is the autoregressive negative log-likelihood over $V_{\mathrm{act}}$, and $\mathcal{L}_{\mathrm{rec}}$ optimizes surgical recommendations. This design fundamentally transforms risk prediction from an isolated classification head into a structural prior that governs generation.

## 3 Experiments

### 3.1 Dataset

To rigorously evaluate our framework, we curate **M$^3$TAVR** (Multimodal, Multi-granularity, Multi-task cohort for TAVR), the largest structural heart disease benchmark to date, comprising 1,482 patients. Each patient record is strictly paired with three modalities: (i) 3D cardiac CT volumes, (ii) 2D echocardiography clips, and (iii) a comprehensive vector of clinical and pathological biomarkers $X_{\mathrm{tab}}$. Furthermore, each case is annotated with a ground-truth procedural risk level $r \in \{l, m, h\}$ and a clinician-authored structured report $Y$ containing diagnostic findings and surgical recommendations (e.g., valve sizing and access routes). To prevent data leakage and ensure robust evaluation, we employ a strict patient-level stratified split: 70% for training, 10% for validation, and 20% for testing. For a randomly sampled subset of the test set ($N = 200$), we additionally obtained expert-annotated region-of-interests (ROIs) for TAVR-critical structures to enable quantitative grounding evaluation.

**Table 1.** Comprehensive benchmarking on the **M$^3$TAVR** test set. Models are grouped into four paradigms. Best results are in **bold**, and second-best are underlined.

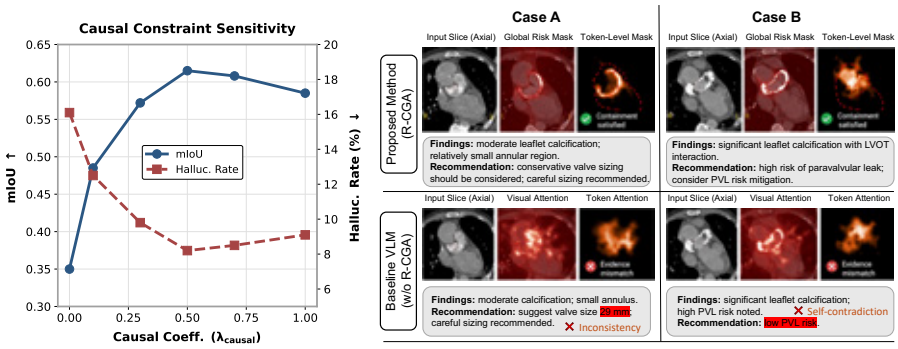
| Model | Risk Prediction | | Report Generation | | | Safety & Grounding | |
|---|---|---|---|---|---|---|---|
| | AUROC ↑ | Macro-F1 ↑ | ROUGE-L ↑ | CIDEr ↑ | ClinEnt-F1 ↑ | Halluc. ↓ | mIoU ↑ |
| *Risk-Only Baselines* | | | | | | | |
| XGBoost [2] | 0.785 | 0.742 | – | – | – | – | – |
| TabNet [1] | 0.791 | 0.750 | – | – | – | – | – |
| Late-Fusion (CT+Echo+Tab) | 0.832 | 0.796 | – | – | – | – | – |
| *Report Generation Baselines* | | | | | | | |
| MCLN-Transformer [3] | – | – | 0.316 | 0.662 | 0.432 | 37.9% | 0.175 |
| R2Gen-Mamba [17] | – | – | 0.329 | 0.684 | 0.458 | 34.6% | 0.210 |
| METransformer [21] | – | – | 0.341 | 0.703 | 0.470 | 33.8% | 0.234 |
| *Open-Source Multimodal Models* | | | | | | | |
| LLaVA-Med-1.5 [8] | 0.804 | 0.773 | 0.388 | 0.821 | 0.614 | 24.2% | 0.454 |
| Qwen3-VL-8B [23] | 0.849 | 0.823 | 0.401 | 0.868 | 0.695 | 17.9% | 0.501 |
| InternVL3.5-8B [20] | 0.861 | 0.836 | 0.409 | 0.889 | 0.724 | 14.7% | 0.516 |
| *Closed-Source Frontier Models* | | | | | | | |
| GPT-4o | 0.868 | 0.844 | 0.412 | 0.901 | 0.756 | 13.8% | – |
| Claude-4 Opus | 0.875 | 0.852 | 0.415 | 0.908 | 0.764 | 12.9% | – |
| Gemini-3 Pro | 0.883 | 0.860 | 0.418 | 0.914 | 0.772 | 11.5% | – |
| *TAVR-VLM (Proposed Framework)* | | | | | | | |
| **R-CGA + Qwen3-VL-8B** | 0.892 | 0.869 | 0.422 | 0.928 | 0.789 | 8.7% | 0.603 |
| **R-CGA + InternVL3.5-8B** | **0.896** | **0.874** | **0.426** | **0.936** | **0.801** | **8.1%** | **0.624** |

### 3.2 Experimental Setup and Evaluation Metrics

To ensure a comprehensive evaluation, we benchmark our framework against four distinct paradigms: risk-only classifiers, medical report generators, open-source MLLMs, and closed-source frontier models, extracting standardized 2D slices for 2D-centric baselines to maintain domain fairness. Our architecture, driven by a 3D ViT for CT extraction and a temporal 2D-CNN for echocardiography, is optimized end-to-end via AdamW across four NVIDIA A100 GPUs with the causal consistency coefficient $\lambda_{\text{causal}}$ empirically set to 0.5. We evaluate the joint modeling capabilities across three critical dimensions: clinical prediction performance (AUROC and Macro-F1), generation fidelity (ROUGE-L [10] and CIDEr [19]). To rigorously penalize diagnostic fabrication, we define Hallucination Rate (HR) as the percentage of generated clinical entities lacking corresponding multimodal evidence, while spatial anchoring precision is quantified via mean Intersection over Union (mIoU) between the constrained token-level mask $\widetilde{M}_t$ and expert-annotated ROIs.

### 3.3 Main Results

Table 1 presents a comprehensive quantitative comparison on the **M$^3$TAVR** test set. By benchmarking against four distinct paradigms, we demonstrate that our proposed R-CGA framework establishes a new state-of-the-art across all evaluated clinical and generation dimensions.




(a) Parameter Sensitivity (b) Qualitative Case Study

**Fig. 2. Comprehensive Model Analysis. (a)** Sensitivity of the causal consistency coefficient $\lambda_{\text{causal}}$ on spatial grounding accuracy and hallucination rate. **(b)** Qualitative comparison between R-CGA and a baseline VLM without risk-conditioned grounding.

*Superiority in Clinical Prediction and Generation.* While current open-source models and cutting-edge closed-source VLMs exhibit formidable general vision-language capabilities, they lack the explicit domain constraints required for high-stakes surgical planning. By integrating our plug-and-play R-CGA module into an 8B-parameter backbone (InternVL3.5-8B), we observe a transformative performance leap. Specifically, our model achieves an AUROC of 0.896 in risk prediction, outperforming the top-performing frontier model, Gemini-3 Pro (0.883). In report generation, our framework surpasses Gemini-3 Pro in CIDEr (0.936 vs. 0.914), indicating a higher consensus with clinical experts in utilizing precise TAVR-specific terminology rather than generic descriptions.

*Breakthrough in Safety and Spatial Grounding.* Diagnostic hallucination remains the most critical bottleneck for MLLM adoption in structural heart interventions. As shown in Table 1, traditional report generation baselines suffer from severe hallucinations ($> 33\%$) and diffuse visual attention (mIoU $< 0.25$). Even the most advanced frontier model, Gemini-3 Pro, yields a hallucination rate of 11.5%, as its autoregressive generation relies heavily on generic memory priors rather than rigorous visual evidence.

In stark contrast, the top-down causal bottleneck in R-CGA reduces hallucinations through risk-conditioned grounding. Enforced by this structural prior, our hallucination rate decreases to 8.1%, while token-to-region grounding precision (mIoU) surges to 0.624. This empirically validates that our model does not merely memorize linguistic distributions; it accurately localizes pathological evidence (e.g., specific calcified leaflets) before emitting the corresponding clinical entities.

### 3.4 Ablation Study

**Module Ablation.** To disentangle the contributions of the R-CGA components, we conduct an ablation study (Table 2). Removing the learnable risk

**Table 2.** Ablation of key R-CGA components on **M$^3$TAVR**.

| Variant | AUROC ↑ | CIDEr ↑ | Halluc. Rate ↓ | mIoU ↑ |
|---|---|---|---|---|
| **Full TAVR-VLM** | **0.896** | **0.936** | **8.1%** | **0.624** |
| w/o Risk Prototypes $W$ | 0.875 | 0.890 | 12.5% | 0.542 |
| w/o Purification ($V_{\text{act}} = V$) | 0.881 | 0.865 | 18.4% | 0.485 |
| w/o Causal Loss ($\lambda_{\text{causal}} = 0$) | 0.885 | 0.872 | 16.1% | 0.350 |
| w/o Stop-Gradient | 0.888 | 0.895 | 14.3% | 0.412 |

prototypes ($W$) degrades the macro bottleneck representation, lowering AUROC to 0.875. Crucially, bypassing the risk-driven visual purification (w/o $V_{\text{act}}$) exposes the language model to unconstrained tokens, triggering an immediate spike in hallucination rate (from 8.1% to 18.4%) as the generator exploits spurious correlations. Disabling the causal consistency loss ($\lambda_{\text{causal}} = 0$) severs the top-down alignment, causing a steep drop in spatial grounding precision (mIoU falls to 0.350). Finally, replacing the stop-gradient mechanism with standard backpropagation (w/o StopGrad) results in degenerate optimization; the model artificially inflates $M_{\text{global}}$ to trivially minimize KL divergence, validating that risk must serve as a strict, unidirectional bottleneck.

**Parameter Sensitivity.** We analyze the sensitivity of the causal consistency coefficient $\lambda_{\text{causal}}$ (Fig. 2a). A clear trade-off emerges: increasing $\lambda_{\text{causal}}$ from 0 to 0.5 drastically improves mIoU (from 0.350 to 0.624) while curbing the diagnostic hallucination rate down to 8.2%. However, overly aggressive penalization ($\lambda_{\text{causal}} > 0.5$) marginally degrades performance, indicating that the language generator becomes too rigidly constrained by the global mask, sacrificing linguistic flexibility. Based on these dynamics, we adopt $\lambda_{\text{causal}} = 0.5$ and a prototype dimension of $d = 512$ for all experiments.

**Qualitative Case Study.** To intuitively understand how R-CGA suppresses hallucinations, we visualize representative cases comparing R-CGA with a baseline VLM without risk-conditioned grounding in Fig. 2b. The baseline model often covers diffuse or clinically irrelevant regions, leading to a mismatch of evidence and unsupported recommendations. In contrast, R-CGA first selects a global risk-relevant region and then constrains token-level grounding to this region through support projection. This mechanism reduces unsupported clinical entities and improves the anatomical consistency between the visual evidence and the generated report statements.

## 4 Conclusion

We presented TAVR-VLM, a pioneering risk-conditioned multimodal framework for hallucination-resistant TAVR report generation. The proposed R-CGA module constructs a model-internal "Risk → Region → Word" structural pathway by

using a causal risk bottleneck to guide anatomical evidence selection and token-level grounding. Unlike unconstrained generation, R-CGA projects clinically significant token attention into a risk-defined support mask and penalizes grounding outside this support, thus mitigating evidence-mismatched clinical statements. Experiments on the $M^3$TAVR dataset demonstrate that TAVR-VLM outperforms both specialized medical models and cutting-edge frontier VLMs in risk prediction, report generation, hallucination reduction, and spatial grounding. Future work will focus on external validation, more detailed clinical endpoint modeling, and broader evaluation across structural heart interventions.

## References

1. Arik, S.Ö., Pfister, T.: Tabnet: Attentive interpretable tabular learning. Proceedings of the AAAI Conference on Artificial Intelligence (2021). https://doi.org/10.1609/aaai.v35i8.16826
2. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2016). https://doi.org/10.1145/2939672.2939785
3. Chen, Z., Song, Y., Chang, T.H., Wan, X.: Generating radiology reports via memory-driven transformer. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1439–1449. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.emnlp-main.112
4. Gheini, M., Ren, X., May, J.: Cross-attention is all you need: Adapting pretrained Transformers for machine translation. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 1754–1765. Association for Computational Linguistics (Nov 2021). https://doi.org/10.18653/v1/2021.emnlp-main.132
5. Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., Liu, T.: A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Trans. Inf. Syst. **43**(2) (Jan 2025). https://doi.org/10.1145/3703155
6. Johnson, A.E.W., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S.: MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Scientific Data (2019). https://doi.org/10.1038/s41597-019-0322-0
7. Leon, M.B., Smith, C.R., Mack, M., Miller, D.C., Moses, J.W., Svensson, L.G., et al.: Transcatheter aortic-valve implantation for aortic stenosis in patients who cannot undergo surgery. New England Journal of Medicine (2010). https://doi.org/10.1056/NEJMoa1008232
8. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: LLaVA-med: Training a large language-and-vision assistant for biomedicine in one day. In: Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2023), https://openreview.net/forum?id=GSuP99u2kR
9. Lilly, S.M., Deshmukh, A.J., Epstein, A.E., Ricciardi, M.J., Shreenivas, S., Velagapudi, P., et al.: 2020 ACC expert consensus decision pathway on management of conduction disturbances in patients undergoing transcatheter aor-

tic valve replacement. Journal of the American College of Cardiology (2020). https://doi.org/10.1016/j.jacc.2020.08.050
10. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Workshop on Text Summarization Branches Out (WAS 2004) (2004)
11. Lu, Z., Li, Y., Tang, F., Jiang, Z., Li, C., Zhou, M., Li, T., Su, J.: Deepgb-tb: A risk-balanced cross-attention gradient-boosted convolutional network for rapid, interpretable tuberculosis screening. Proceedings of the AAAI Conference on Artificial Intelligence **40**(46), 38989–38997 (2026). https://doi.org/10.1609/aaai.v40i46.41245
12. Lu, Z., Su, J.: Hierrisk: A hierarchical framework for suicide risk prediction on social media. In: 2025 IEEE International Conference on Big Data (BigData). pp. 8169–8174 (2025). https://doi.org/10.1109/BigData66926.2025.11402629
13. Lu, Z., Su, J.: Dialectic-med: Mitigating diagnostic hallucinations via counterfactual adversarial multi-agent debate (2026), https://arxiv.org/abs/2604.11258
14. Lu, Z., Xu, S., Li, Y., Su, J., Tang, T.: Causal-sam-llm: Large language models as causal reasoners for robust medical segmentation. In: ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 6246–6250 (2026). https://doi.org/10.1109/ICASSP55912.2026.11460869
15. Lu, Z., Xu, S., Yan, K., Cai, X., Zhang, C., Li, Y., Stefanidis, A., Nguyen, A., Su, J.: Skinclip-vl: Consistency-aware vision-language learning for multimodal skin cancer diagnosis (2026), https://arxiv.org/abs/2603.21010
16. Mack, M.J., Leon, M.B., Thourani, V.H., Makkar, R.R., Kodali, S.K., Russo, M., et al.: Transcatheter aortic-valve replacement with a balloon-expandable valve in low-risk patients. New England Journal of Medicine (2019). https://doi.org/10.1056/NEJMoa1814052
17. Sun, Y., Lee, Y.Z., Woodard, G.A., Zhu, H., Lian, C., Liu, M.: R2gen-mamba: A selective state space model for radiology report generation. In: 2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI). pp. 1–4 (2025). https://doi.org/10.1109/ISBI60581.2025.10980814
18. VARC-3 Writing Committee, Généreux, P., Piazza, N., Alu, M.C., Nazif, T., Hahn, R.T., et al.: Valve academic research consortium 3: Updated endpoint definitions for aortic valve clinical research. Journal of the American College of Cardiology (2021). https://doi.org/10.1016/j.jacc.2021.02.038
19. Vedantam, R., Zitnick, C.L., Parikh, D.: Cider: Consensus-based image description evaluation. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4566–4575 (2015). https://doi.org/10.1109/CVPR.2015.7299087
20. Wang, W., et al.: Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency (2025), https://arxiv.org/abs/2508.18265
21. Wang, Z., Liu, L., Wang, L., Zhou, L.: Metransformer: Radiology report generation by transformer with multiple learnable expert tokens. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11558–11567. IEEE Computer Society (2023). https://doi.org/10.1109/CVPR52729.2023.01112
22. Xue, C., Liu, Y., Zhou, M., Su, J., Lu, Z.: Semantic-topological graph reasoning for language-guided pulmonary screening (2026), https://arxiv.org/abs/2604.05620
23. Yang, A., et al.: Qwen3 technical report (2025), https://arxiv.org/abs/2505.09388