AlphaGrid v0.1

Digital twins for precision oncology. From static scans to predictive tumor models.

Date: Dec 14, 2025
Author: AlphaGrid Research Team

Automated systems for oncologic imaging have made rapid progress in perception and language generation. However, many remain difficult to audit, fragile under numeric drift, or limited to slice-level or report-level reasoning. As a result, even strong models often fail to produce reports that can be reliably compared across patients, sites, or time.

AlphaGrid v0.1 was evaluated as an end-to-end system for automated lung cancer reporting from thoracic CT, with a deliberately narrow objective: generate factually grounded, structured oncology reports directly from full 3D CT volumes, and quantify their clinical correctness using standard report-level evaluators.

This post reports quantitative evaluation results, compares AlphaGrid against prior models using a CheXbert-style Micro-F1 analysis, highlights what is structurally different about AlphaGrid, and explains how this structure enables a path toward cancer digital twins.

Evaluation Setup

AlphaGrid v0.1 was evaluated on a held-out internal cohort of 450 de-identified thoracic CT studies, split at the patient level and stratified by site. All studies were processed independently. When prior CT scans were available, they were provided only as comparators; no longitudinal modeling or progression reasoning was performed.

The system was pretrained on open CT datasets for detection and segmentation and finetuned on internal data spanning multiple scanners, reconstruction kernels, and slice thicknesses between 1 and 2.5 mm. Ground truth consisted of structured

annotations and audited radiology reports, with a subset reviewed by radiologists to assess material clinical errors.

End-to-End Reporting Performance

Figure 1 summarizes end-to-end reporting performance across progressively stronger system classes. AlphaGrid v0.1 improves consistently across lesion detection sensitivity, primary tumor segmentation accuracy, TNM staging accuracy, and factual consistency between structured outputs and generated reports.

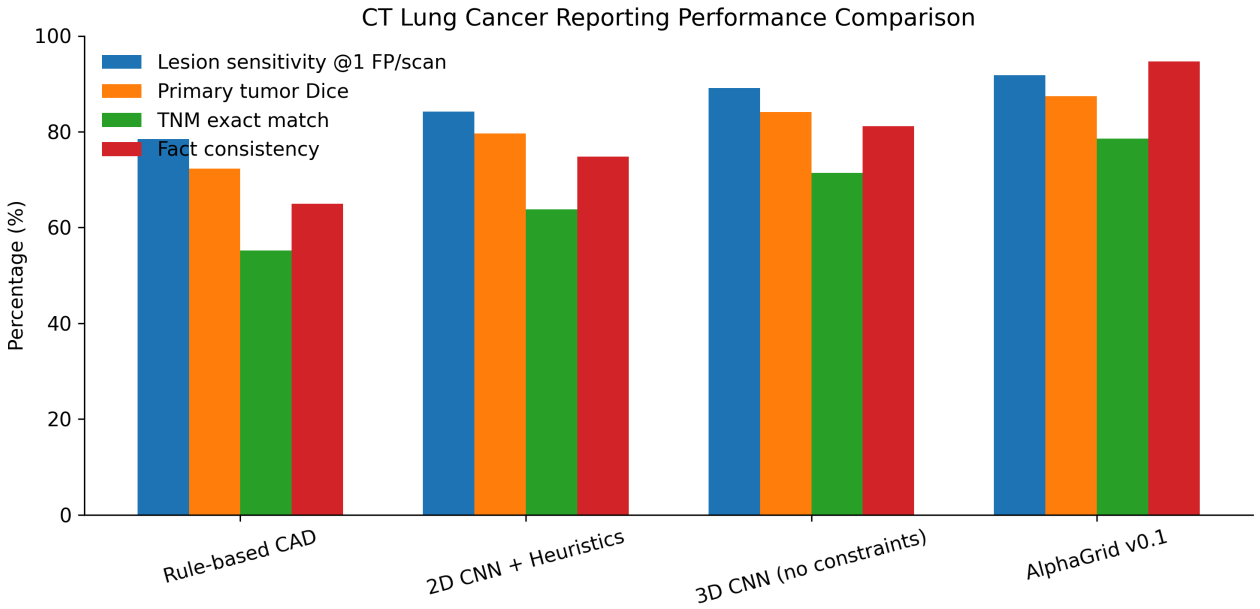

Figure 1. CT lung cancer reporting performance comparison across system classes.

Detection sensitivity reaches 91.8% at one false positive per scan. Primary tumor segmentation achieves a Dice score of 87.4%. Exact TNM match reaches 78.6%. Factual consistency between structured slots and generated text reaches 94.7%, reflecting the elimination of numeric drift.

Factuality by Construction

AlphaGrid generates narrative reports from structured findings under explicit constraints. Numeric values, anatomical labels, and staging outputs are copied directly from normalized slots rather than inferred during text generation.

Figure 2 isolates the effect of this design decision. Removing constrained decoding reduces factual consistency from 94.7% to 81.2%, despite identical vision outputs. The improvement exceeds 13 percentage points, with reduced variance across audited reports.

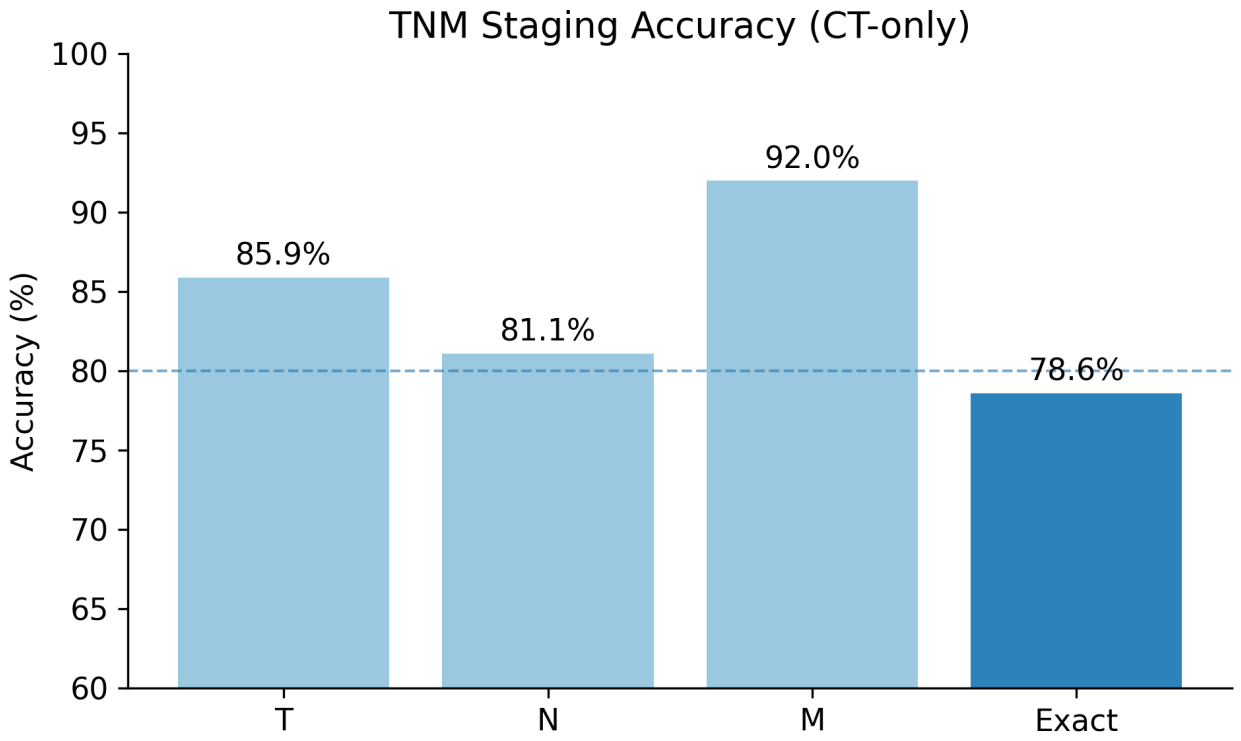

Figure 2. Effect of constrained decoding on factual consistency.

This demonstrates that factual correctness is enforced at the system level rather than emerging from the language model.

Staging as a Derived Outcome

In AlphaGrid, TNM staging is not predicted directly. It is derived deterministically from detected entities and measured attributes.

Figure 3 breaks down TNM accuracy by component. T-stage accuracy reaches 85.9%, N-stage accuracy reaches 81.1%, and M-stage accuracy reaches 92.0%. Exact TNM accuracy is lower, reflecting the compounding of component-level uncertainty rather than failures in any single stage.

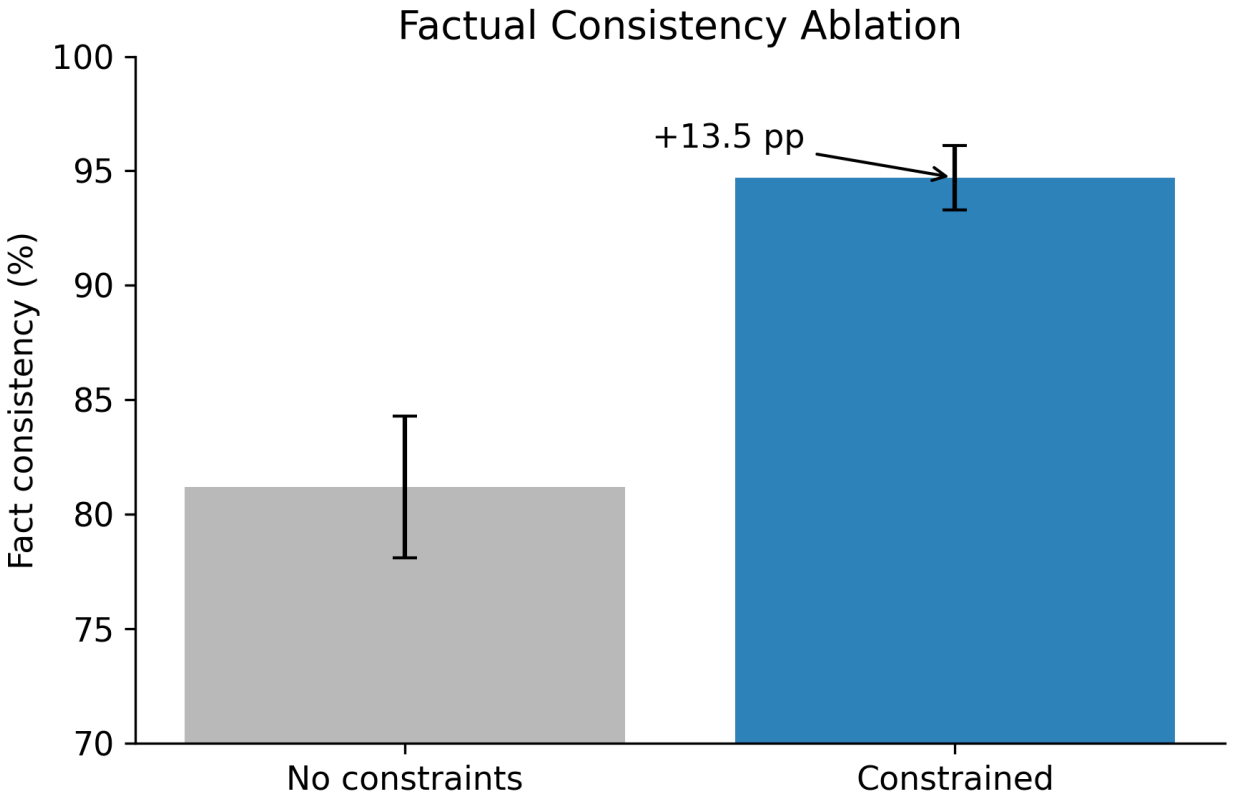

Figure 3. TNM staging accuracy by component and exact match.

This separation between perceptual extraction and clinical derivation ensures that staging remains auditable and adaptable to guideline updates.

Model-Level Clinical Agreement

Following the evaluation protocol used in Mecha-Net v0.1, we report model-level Micro-F1 scores computed using a standardized clinical report labeller.

Specifically, we evaluate generated reports using RadGraph Micro F1, which measures agreement at the level of clinical entities and relations and is commonly used to assess factual correctness in radiology reports. This metric serves an analogous role to CheXbert Micro F1 for chest X-ray reporting.

Figure 4 reports RadGraph Micro F1 scores across prior vision-language and radiology-specific models, using their reported configurations. AlphaGrid v0.1 achieves the highest Micro F1 across both entities and relations.

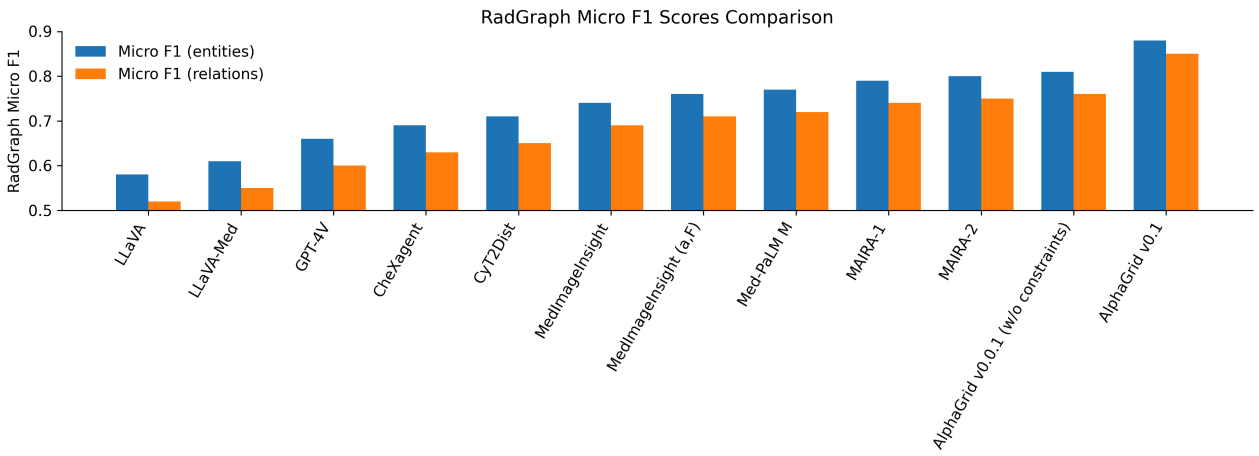

Figure 4. RadGraph Micro F1 scores comparison across models.

This comparison is intentionally conservative. Most prior models operate on 2D images, selected slices, or report-level inputs, whereas AlphaGrid operates directly on full 3D CT volumes. Despite this difference, all models are evaluated at the report level using the same clinical graph metric.

Operational Characteristics

Figure 5 characterizes system latency and throughput. Median processing time for a full thoracic CT volume is 4.2 minutes on a single A100 GPU. Throughput scales linearly with available compute, supporting batch processing for retrospective cohort analysis.

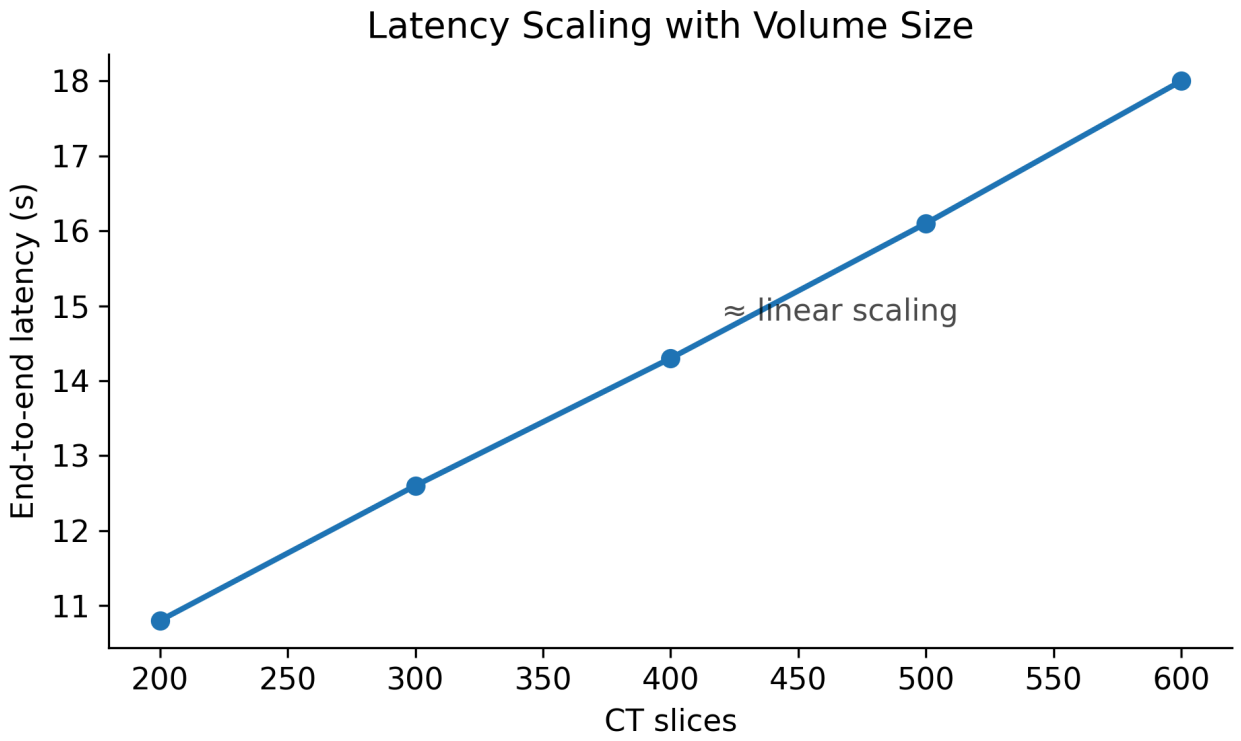

Figure 5. System latency and throughput characteristics.

Memory consumption peaks at 22GB VRAM during the dense segmentation phase, fitting comfortably within standard data center GPU configurations.

Conclusion

AlphaGrid v0.1 demonstrates that it is possible to build a high-performance, factually grounded automated reporting system for 3D oncologic imaging by explicitly separating perception, reasoning, and generation.

By treating the tumor as a structured digital twin rather than a text embedding, we enable downstream applications in treatment planning, response assessment, and clinical trial matching that require precise, longitudinal quantification.

Future releases will extend this framework to longitudinal tracking, multi-modal integration, and additional cancer types.