TITAN: A Multimodal Whole-Slide Foundation Model for Computational Pathology

Authors: Tong Ding, Sophia J. Wagner, Andrew H. Song, Richard J. Chen, Ming Y. Lu, Andrew Zhang, Anurag J. Vaidya, Guillaume Jaume, Muhammad Shaban, Ahrong Kim, Drew F. K. Williamson, Harry Robertson, Bowen Chen, Cristina Almagro-Pérez, Paul Doucet, Sharifa Sahai, Chengkuan Chen, Christina S. Chen, Daisuke Komura, Akihiro Kawabe, Mieko Ochi, Shinya Sato, Tomoyuki Yokose, Yohei Miyagi, Shumpei Ishikawa, Georg Gerber, Tingying Peng, Long Phi Le, Faisal Mahmood

Year: 2025
Venue: Nature Medicine

1. Classification

- **Domain Category:**
  - Medical Vision FM + Medical VLM / MLLM / MMFM
  - TITAN is a whole-slide histopathology foundation model that combines vision-only pretraining with vision–language alignment for pathology reports and synthetic captions.
- **FM Usage Type:**
  - Core FM development + Multimodal FM or cross-modal integration
- **Key Modalities:**
  - Whole-slide histopathology images (WSIs) across ≈20 organ types.
  - Pathology reports (free-text, slide-level).
  - Synthetic fine-grained region-of-interest (ROI) captions generated by a multimodal pathology copilot (PathChat).

2. Executive Summary

TITAN (Transformer-based pathology Image and Text Alignment Network) is a slide-level foundation model for pathology designed to transform gigapixel whole-slide images into general-purpose feature representations that support diagnosis, prognosis, retrieval, and report generation. Instead of working at the level of raw pixels, TITAN builds on pre-extracted patch embeddings from powerful histology encoders, then scales self-supervised learning (SSL) to entire slides using a vision transformer with long-context positional encodings. The model is pretrained in three stages: vision-only self-supervision on hundreds of thousands of WSIs; vision–language alignment using synthetic ROI-level captions; and slide-level alignment with pathology reports. This yields TITANV (vision-only) and full TITAN (vision–language), which are evaluated across slide classification, biomarker prediction, survival analysis, rare cancer retrieval, cross-modal slide–report retrieval, and zero-shot report generation. TITAN consistently outperforms prior ROI-based and slide-level foundation models across linear probing, few-shot, and zero-shot settings, especially in low-data clinical scenarios. For a new grad student, TITAN provides a clear blueprint for scaling from patch encoders to slide-level multimodal FMs in pathology.

3. Problem Setup and Motivation

- **Scientific / practical problem:**
    - Learn **slide-level representations** of histopathology WSIs that:
        - Capture rich tissue morphology at multiple spatial scales.
        - Support a wide range of downstream tasks (subtyping, biomarker prediction, prognosis, retrieval, report generation).
        - Work well even in **low-data and rare disease** regimes.
- **Why this is hard:**
    - **Gigapixel scale and long context:**
        - WSIs can contain >10^4 patch embeddings; naïvely applying transformers is computationally prohibitive.
    - **Limited labeled cohorts:**
        - Clinical datasets for specific cancers or biomarkers are small and heterogeneous, especially for rare conditions.
    - **Patch vs slide gap:**
        - Many existing FMs operate on small ROIs; aggregating patch features into clinically meaningful slide-level signals is non-trivial.
    - **Multimodal supervision:**
        - Pathology reports and textual descriptions encode rich semantics but are noisy and unstructured; exploiting them at scale is challenging.

  - **Generalization and retrieval:**
    - Models must generalize across organs, stains, scanner types, and institutions, and support tasks like rare cancer retrieval where labeled examples are extremely sparse.

4. Data and Modalities

- **Pretraining data (Mass-340K):**
  - ≈335,645 WSIs across 20 organ types.
  - 182,862 human pathology reports at slide level.
  - Diverse stains, tissue types, and scanners to maximize coverage of histopathology morphologies.
- **Synthetic caption data:**
  - 423,122 synthetic ROI-level captions generated from 8k×8k pixel regions using PathChat, a multimodal pathology copilot.
  - Each caption describes fine-grained morphology within an ROI (e.g., cell types, tissue organization).
- **Downstream benchmarks (representative):**
  - Cancer subtyping and grading across multiple tumor types.
  - Molecular biomarker prediction (e.g., mutation status, molecular subtypes).
  - Survival prediction and outcome prognosis.
  - **Rare cancer retrieval:** retrieve similar slides for diagnostically challenging WSIs.
  - Cross-modal retrieval (slide ⍰ report).
  - Zero-shot and few-shot slide classification guided by textual prompts.
- **Preprocessing / representation:**
  - WSIs are divided into 512×512 patches at 20× magnification, and each patch is encoded into a 768-dimensional feature using a

strong patch encoder (CONCH v1.5).

- Patch features are arranged into a 2D grid reflecting spatial layout, then cropped into global and local views for SSL.
- Pathology reports and synthetic captions are tokenized and embedded for vision–language alignment.

5. Model / Foundation Model

- **Model Type:**
    - Slide-level Vision Transformer (ViT) foundation model with multimodal vision–language pretraining.
- **Is it a new FM or an existing one?**
    - TITAN is a **new whole-slide foundation model**, though it builds on existing patch encoders (e.g., CONCH) and SSL techniques (iBOT, CoCa-style alignment).
- **Key components and innovations:**

Aspect	Details
Backbone	ViT-style transformer operating on patch-feature tokens
Input tokens	2D grid of patch embeddings (from CONCH) plus [CLS] / slide tokens
Vision-only pretraining	iBOT-style masked prediction on WSI feature grids (TITANV)

Aspect	Details
ROI-level alignment	Contrastive alignment with synthetic ROI captions (PathChat)
Slide-level alignment	Contrastive / CoCa-style alignment with pathology reports
Positional encoding	Long-range encodings (e.g., ALiBi-style) adapted to large 2D grids

- **Training setup (three stages):**
  - **Stage 1 – Vision-only SSL (TITANV):**
    - Perform iBOT pretraining on region crops (16×16 token grids) and their multi-scale views (global 14×14 and local 6×6 crops).
    - Learns slide-level representations that aggregate patch-level morphologies.
  - **Stage 2 – ROI-level vision–language pretraining:**
    - Align 423k ROI crops (8k×8k regions) with synthetic captions from PathChat.
    - Encourages TITAN to associate specific morphological patterns with textual descriptions.
  - **Stage 3 – Slide-level vision–language pretraining:**
    - Align 183k WSIs with their corresponding pathology reports, enabling slide-level semantic understanding and cross-modal retrieval.
  - After pretraining, TITAN can be used for linear probing, few-shot fine-tuning, zero-shot classification via text prompts, and report generation.

6. Multimodal / Integration Aspects (If Applicable)

- **Modalities integrated:**
  - Histopathology WSIs and free-text pathology reports, plus synthetic ROI-level captions.
- **How integration works:**
  - **Vision-only backbone:**
    - TITANV is trained with SSL on slide-level patch features to learn rich visual embeddings.
  - **ROI-level vision–language alignment:**
    - A contrastive or CoCa-style loss aligns ROI embeddings with synthetic PathChat captions, injecting fine-grained morphological semantics.
  - **Slide-level vision–language alignment:**
    - Slide embeddings are aligned with pathology report text, enabling cross-modal retrieval and zero-shot, text-prompted classification.
- **New capabilities enabled:**
  - **Zero-shot and few-shot slide classification** using textual prompts describing subtypes or biomarkers.
  - **Rare cancer retrieval:** find clinically similar slides based on TITAN embeddings, even with minimal labeled examples.
  - **Cross-modal search:** slide-to-report and report-to-slide retrieval for case exploration and education.
  - **Report generation:** generate slide-level pathology reports conditioned on WSI embeddings and language decoders.

7. Experiments and Results

- **Tasks / benchmarks:**
  - Slide-level cancer subtyping and grading across multiple public and internal cohorts.
  - Molecular prediction (e.g., mutation and expression surrogates) from WSIs.
  - Survival prediction and risk stratification.
  - Rare cancer and challenging-case retrieval.
  - Cross-modal slide–report retrieval and zero-shot classification using textual prompts.
- **Baselines:**
  - ROI-based patch encoders combined with slide aggregators (MIL and attention-pooling models).
  - Prior slide-level foundation models trained with vision-only SSL or smaller multimodal datasets.
  - Task-specific supervised models trained on individual cohorts.
- **Key findings (trends):**
  - TITANV (vision-only) already outperforms prior slide-level models and ROI-aggregator baselines on many slide classification and biomarker tasks.
  - Full TITAN (after multimodal pretraining) further improves performance, particularly in **low-data and few-shot** settings.
  - TITAN shows strong **rare cancer retrieval** performance, retrieving pathologically similar slides that can assist in challenging diagnoses.
  - Vision–language pretraining with synthetic ROI captions and reports enables **zero-shot text-guided classification** and cross-modal retrieval that prior slide FMs cannot match.

8. Strengths, Limitations, and Open Questions

Strengths:

- First large-scale **multimodal whole-slide foundation model** that unifies vision-only SSL and vision–language alignment across ROI and slide levels.
- Demonstrates that building on strong patch encoders and scaling to slide level yields **state-of-the-art performance** across many pathology tasks.
- Synthetic ROI captions from PathChat provide a practical way to incorporate fine-grained morphological supervision at scale.
- Extensive evaluations across tasks, organs, and settings (linear probing, few-shot, zero-shot) show TITAN's breadth and robustness.

Limitations:

- Relies on a large, internal Mass-340K dataset and synthetic captions that are not fully public, limiting reproducibility.
- Synthetic captions, while powerful, may encode **biases and failure modes** of the PathChat generator.
- Focuses primarily on histopathology WSIs; other modalities (radiology, multi-omics, clinical text beyond reports) are not integrated.
- Training at this scale requires substantial compute and storage, making it difficult for smaller groups to replicate or extend.

Open Questions and Future Directions:

1. How can TITAN-style slide-level FMs be **extended to multimodal clinical contexts**, integrating WSIs with genomics, radiology, and EHR data?
2. What are robust methods for **validating and correcting synthetic captions**, ensuring that vision–language supervision does not propagate hallucinations into the slide encoder?

3. Can more efficient architectures (e.g., sparse attention, hierarchical transformers) reduce the cost of handling giga-pixel WSIs without losing performance?
4. How should we design evaluation protocols and human-in-the-loop workflows for **rare cancer retrieval**, where errors may have significant diagnostic consequences?
5. Could TITAN representations support **interactive, region-grounded explanations** that show which slide regions drive predictions or retrieved cases?

9. Context and Broader Impact

- **Position in the landscape:**
  - TITAN is to **pathology WSIs** what CLIP-like and CoCa-style models are to natural images and text: a general-purpose slide representation that supports many downstream tasks and multimodal interactions.
  - It extends the trend of patch-level pathology FMs to the slide level, helping bridge the gap between detailed morphology and patient-level clinical endpoints.
- **Relation to well-known ideas:**
  - Combines **iBOT-style masked image modeling**, **patch-encoder distillation**, and **vision–language contrastive pretraining** in a three-stage pipeline.
  - Conceptually similar to **GigaPath** and other large-scale pathology FMs but emphasizes multimodal, slide-level pretraining and rare cancer retrieval.
- **Why this paper is a useful reference:**
  - Provides a detailed design for scaling from patch encoders to slide-level transformers and for incorporating synthetic and real textual supervision.

    - For a grad student, TITAN is an archetype for building high-capacity medical vision FMs and integrating them into multimodal medical AI systems.

10. Key Takeaways (Bullet Summary)

- **Problem:**
    - Existing pathology FMs mostly work at the patch level and lack robust, multimodal slide-level representations, limiting performance in patient-level prediction and rare disease scenarios.
- **Method / model:**
    - TITAN is a ViT-based slide foundation model trained in three stages: large-scale vision-only SSL on WSI patch features, ROI-level alignment with synthetic PathChat captions, and slide-level alignment with pathology reports.
    - Operates entirely in the patch-embedding space, with long-range positional encodings and multi-scale cropping to handle giga-pixel WSIs.
- **Results:**
    - Outperforms ROI-based and prior slide-level FMs on cancer subtyping, biomarker prediction, prognosis, and retrieval, especially in low-data and few-shot regimes.
    - Enables zero-shot text-guided classification, cross-modal slide–report retrieval, and pathology report generation.
- **Why it matters:**
    - Establishes a strong template for **multimodal slide-level foundation models**, bringing pathology closer to the capabilities seen in natural-image FMs and VLMs.
    - Opens pathways for rare disease support, better education and retrieval tools, and future integration with other medical modalities.