Research Statement

Knowledge-Grounded and Efficient Multimodal Learning Systems

Multimodal foundation models achieve strong empirical performance, yet their learning paradigm remains largely static. Knowledge is encoded implicitly in parameters, visual reasoning is often mediated through language backbones, and integrating structured signals such as knowledge graphs can destabilize optimization or increase memory complexity. As a result, structured reasoning and adaptation to evolving information require expensive retraining rather than principled updates. My research develops knowledge-grounded and computationally efficient multimodal systems that treat structured knowledge as an explicit component of the learning objective. By designing structured learning frameworks and well-conditioned optimization methods, I enable robust integration, improved convergence behavior, and controlled model adaptation. This research agenda is organized around three interconnected directions::

- ○ **Structured Multimodal Representation (§1) :** I design structured learning and objective formulations that induce relational and multimodal structure directly on embedding space. By casting knowledge integration as a constrained optimization problem, I develop robust mechanisms for incorporating structured signals into multimodal models.
- ○ **Objective Conditioning and Optimization Dynamics (§2) :** I analyze how curvature and implicit differentiation govern convergence, robustness, and memory complexity. Through well-conditioned objective formulations and fixed-point network operators, I improve first-order optimization dynamics under realistic computational constraints.
- ○ **Constrained Adaptation in Evolving Models (§3) :** I formulate selective model updates as constrained optimization problems with convergence guarantees and controllable trade-offs, enabling model adaptation under distribution shift without full retraining.

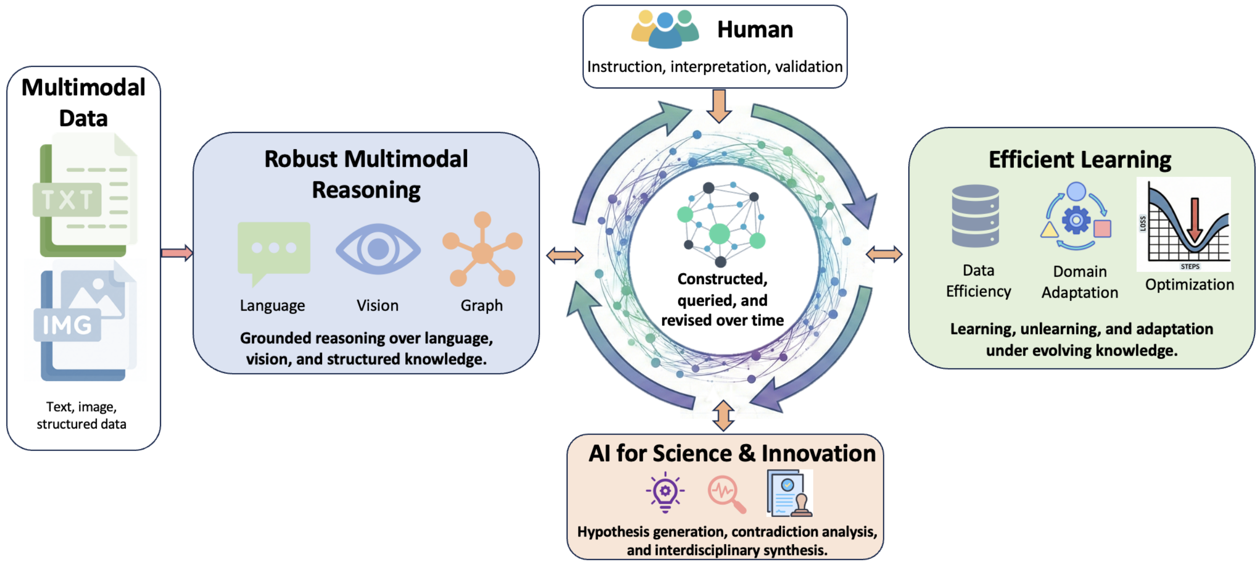
Multimodal Data
TXT
IMG
Text, image, structured data
Robust Multimodal Reasoning
Language
Vision
Graph
Grounded reasoning over language, vision, and structured knowledge.
Human
Instruction, interpretation, validation
Constructed, queried, and revised over time
Efficient Learning
Data Efficiency
Domain Adaptation
LOSS
STEPS
Optimization
Learning, unlearning, and adaptation under evolving knowledge.
AI for Science & Innovation
Hypothesis generation, contradiction analysis, and interdisciplinary synthesis.

Figure 1. Research vision. Structured learning shapes multimodal representations, conditioned objectives stabilize learning dynamics, and constrained adaptation enables selective model revision as knowledge evolves. Together, these components form a unified framework for robust, efficient, and scalable multimodal reasoning in dynamic environments.

1. Structured Multimodal Representation

Advanced multimodal models rely on high-capacity embeddings, yet the geometry of these representations is typically shaped implicitly by data statistics and large-scale pretraining. When structured knowledge is introduced, such as relational graphs, long-tailed class hierarchies, and multimodal alignment signals, it is often injected heuristically, leading to instability or misalignment in embedding space. My work addresses this limitation by treating structured knowledge as an explicit structural constraint on representation space.

Relational Objectives and Contextualized Structural Embeddings. In early work on ontology matching [3, 4], I developed self-supervised objectives that move beyond local triple supervision to incorporate structural path dependencies, inducing globally consistent relational geometry in embedding space. This formulation captures structural context without requiring labeled alignment. Extending this perspective to multimodal systems [7], I designed class-aware contrastive objectives that explicitly reweight alignment under long-tailed distributions, reshaping embedding alignment to reflect structural imbalance. Across these settings, structure is encoded through objective transformations that alter the structure of representation space rather than post-hoc retrieval or external filtering.

Implicit and Diffusion-Based Operators for Robust Multimodal Integration. In knowledge-grounded multimodal models, directly incorporating external signals can introduce high-variance perturbations that destabilize optimization. To address this [5], I introduced an implicit fixed-point operator that filters structured signals through outlier detection, enabling stable integration while supporting Jacobian-free differentiation and constant memory backpropagation. Complementarily, I developed diffusion-based operators [6] that transform multimodal embeddings prior to downstream reasoning, inducing structured smoothness in feature space and enabling recursive graph-based partitioning. These methods unify relational constraints, multimodal alignment, and structured transformation within a unified structural framework.

2. Optimization for Scalable Multimodal Learning

Despite advances in representation design, the stability and efficiency of multimodal systems depend critically on the conditioning and curvature of their learning objectives. In large-scale settings, poorly conditioned loss landscape amplify gradient variance, slow convergence, and increase sensitivity to quantization and structured perturbations. My work investigates how conditioning and implicit differentiation shape optimization dynamics, with the goal of designing objectives whose stability properties are intrinsic rather than emergent from scale.

Well-Conditioned Objective Design. In Rooted Logistic Objectives [9], I introduced a loss formulation that modifies the curvature profile of standard cross-entropy by improving its condition number. I establish strict convexity properties and analyze how improved conditioning accelerates first-order convergence while enhancing robustness under quantization and pruning. This perspective reframes loss design as a problem of objective conditioning: by reshaping curvature, one can systematically improve optimization behavior without increasing model capacity.

Implicit Differentiation and Memory-Efficient Optimization. Structured multimodal models often incorporate auxiliary operators whose gradients require storing full computational trajectories. To address this, I developed implicit fixed-point formulations [5] that enable

Jacobian-free differentiation, reducing memory complexity from trajectory dependent backpropagation to constant memory updates. This framework connects conditioning, equilibrium constraints, and resource efficiency within a unified optimization perspective.

3. Constrained Adaptation in Evolving Models

Learning systems deployed in non-stationary environments must accommodate continual change: new data distributions emerge, structured knowledge evolves, and domain constraints shift over time. However, conventional retraining treats adaptation as full re-optimization, which disregards previously learned structure and incurs substantial computational cost. In contrast, I formulate model adaptation as a constrained optimization problem, enabling selective parameter updates and structured behavioral control while preserving global consistency and stability.

Bilevel Optimization for Selective Model Revision. I develop a bilevel optimization framework [10] in which an upper-level objective enforces targeted decision boundary adjustments and a lower-level objective preserves retained knowledge. This formulation exposes the forgetting–retention tradeoff as an explicit regularization parameter and admits iterative first-order algorithms with convergence guarantees. By casting model revision as structured optimization, parameter updates become principled and controllable rather than heuristic re-training or fine-tuning.

Structured Adaptation in Generative and Multimodal Systems. Adaptation extends beyond parameter removal to systems whose representations and outputs must remain coherent under evolving semantic and relational constraints. In structure-guided response generation [1, 2], I integrate explicit semantic signals, such as annotated causal spans and figurative language, into the training objective and prompting interface, treating these signals as structured constraints that regulate decoding dynamics. Complementarily, in large-scale multimodal ecosystems such as IMPACT [11] and TRIZBench [8], I model technological artifacts as evolving multimodal knowledge whose representations must integrate new datapoints while preserving higher-order structural consistency. In ongoing work, I further investigate how multimodal representations induce relational networks over scientific and technological artifacts, enabling the analysis of structural evolution, clustering dynamics, and knowledge propagation at scale.

Future Research

My future research advances knowledge-grounded and computationally efficient multimodal systems by extending the three pillars of representation geometry, objective conditioning, and constrained adaptation toward large-scale scientific discovery. The overarching goal is to design AI systems that not only learn from multimodal data, but also evolve with structured knowledge while maintaining robustness, efficiency, and interpretability.

Scalable Structured Knowledge Infrastructure. I will develop scalable structured knowledge infrastructures that enable foundation models to interact with evolving multimodal information in a principled and transferable manner. Rather than treating knowledge graphs as static external databases, I aim to design frameworks that construct, update, and validate structured representations from heterogeneous multimodal data across domains. This direction investigates how structured knowledge layers can serve as adaptive memory systems for foundation models, supporting retrieval, reasoning, and revision without full retraining.

Optimization Dynamics and Robust Learning. I will advance a unified theory of learning dynamics and robustness in structured multimodal systems. Specifically, I will investigate how momentum-based optimizers and low-precision training interact with structured objectives and multimodal constraints. A central goal is to understand how optimization and training dynamics co-evolve under quantization, pruning, and continual updates. By designing optimization methods whose stability and efficiency are intrinsic to their formulation, I aim to develop scalable multimodal learning systems that remain robust under realistic computational and deployment constraints.

Principled Adaptation and Human–AI Co-Discovery. I will investigate constrained adaptation mechanisms that enable multimodal systems to function as evolving collaborators in scientific discovery. Specifically, models will update selectively through structured revision that preserves consistency while integrating new evidence. In domains such as technological innovation analysis [8] and scientific hypothesis generation, this framework enables AI systems to organize multimodal artifacts, highlight structural patterns, and propose candidate connections while remaining grounded in interpretable knowledge representations. Human expertise remains central, with adaptation mechanisms designed to support interpretation and iterative refinement.

Taken together, these directions define a coherent research agenda centered on structured representation, conditioned optimization, and constrained evolution of multimodal systems. My future work aims to build AI systems that are not merely large, but structurally grounded, computationally efficient, and capable of evolving with scientific knowledge.

References

[1] Lee, G., **Wang, Z.**, Ravi, S. N., and Parde, N. EmpatheticFIG at WASSA 2024 empathy and personality shared task: Predicting empathy and emotion in conversations with figurative language. In *WASSA 2024 Shared-Task: Empathy Detection and Emotion Classification* (2024).

[2] Lee, G., **Wang, Z.**, Ravi, S. N., and Parde, N. From heart to words: Generating empathetic responses via integrated figurative language and semantic context signals. In *The 63rd Annual Meeting of the Association for Computational Linguistics* (2025).

[3] **Wang, Z.** Amd results for oaei 2022. In *The 17th International Workshop on Ontology Matching* (2022).

[4] **Wang, Z.** Contextualized structural self-supervised learning for ontology matching. In *The 18th International Workshop on Ontology Matching* (2023).

[5] **Wang, Z.**, Medya, S., and Ravi, S. N. Implicit differentiable outlier detection enable robust deep multimodal analysis. In *Thirty-seventh Conference on Neural Information Processing Systems* (2023).

[6] **Wang, Z.**, Mishra, H., and Ravi, S. N. Improving training-free open-vocabulary segmentation using diffused cuts. In *Under review.*

[7] **Wang, Z.**, Shomee, H. H., Ravi, S. N., and Medya, S. DesignCLIP: Multimodal learning with CLIP for design patent understanding. In *The 2025 Conference on Empirical Methods in Natural Language Processing* (2025).

[8] **Wang, Z.**, and Uzzi, B. Inventive problem solving with llms: A benchmark for triz reasoning. In *Under review.*

[9] **Wang, Z.**, Veluswami, P. R., Mishra, H., and Ravi, S. N. Optimizing neural network training and quantization with rooted logistic objectives. In *The 28th International Conference on Artificial Intelligence and Statistics* (2025).

[10] Nahass, G., **Wang, Z.**, Rashidisabet, H., Won Hwa Kim, S. H., Peterson, J. C., Yazdanpanah, G., Purnell, C. A., Setabutr, P., Tran, A. Q., Yi, D., and Ravi, S. N. Targeted unlearning using perturbed sign gradient methods with applications on medical images. *Transactions on Machine Learning Research* (2025).

[11] Shomee, H. H., **Wang, Z.**, Ravi, S. N., and Medya, S. IMPACT: A large-scale integrated multimodal patent analysis and creation dataset for design patents. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track* (2024).