Thinking in Space with Transformer^2: How MLLMs Can Leverage SVF and Two-Pass Inference to Improve Spatial Reasoning

Ayush Bodade
ayushbodade1@gmail.com

18 January 2025

Contents

Introduction

The demand for advanced spatial reasoning and multimodal understanding has driven the development of methods for MLLMs to think in space. These systems must process and interact with spatial environments across text, vision, and other modalities. However, the diverse requirements of these tasks pose challenges in scalability, efficiency, and adaptability.

This document details the integration of the Transformer^2 self-adaptive framework with *Thinking in Space"* to optimize performance in spatial reasoning tasks. Transformer^2 incorporates Singular Value Fine-tuning (SVF) and modular adaptation strategies to enable scalable, efficient, and adaptive enhancements for multimodal models. By employing SVF and 2-pass inference, MLLMs can efficiently think in space and adapt to varying spatial reasoning tasks while ensuring computational scalability.

Connections Between Thinking in Space and 3D Systems

Perception and 3D Reconstruction:

Connection: Spatial reasoning begins with accurate perception, forming the foundation for 3D reconstruction. Both rely on processing raw sensory data (e.g., video, depth, RGB-D) to build representations of objects and their spatial relationships.

Details:

- In *"Thinking in Space"*, Multimodal Large Language Models (MLLMs) perceive objects in relation to one another, estimating distances, directions, and sizes from video sequences.
- In 3D reconstruction, this corresponds to capturing geometry, texture, and object positions to reconstruct a digital twin of a scene.

Transformer^2 Integration:

- Use Transformer^2's *expert vectors* trained on tasks like object recognition and spatial layout estimation to enhance perception and 3D reconstruction.
- Employ the two-pass inference to resolve ambiguities in noisy or incomplete scans by dynamically refining object relationships.
- Combine SVF with a feedback mechanism to infer occluded or hidden parts of objects (e.g., estimating a table leg hidden by a chair).

Future Potential: Integrating Transformer^2 could reduce dependency on perfect visual input, improving reconstruction in cluttered or complex environments.

Egocentric-Allocentric Transformation for Global Spatial Consistency:

Connection: Transforming egocentric (first-person) views into allocentric (global, map-like) perspectives is crucial in spatial reasoning and 3D reconstruction.

Details:

- Egocentric-allocentric transformations allow models to reason about spatial layouts irrespective of camera position or movement.
- In 3D reconstruction, aligning frames or point clouds from moving cameras requires global consistency across viewpoints.

Transformer^2 Integration:

- Train *expert vectors* for specific spatial alignment tasks (e.g., stitching frames into allocentric maps).
- Use Transformer^2's two-pass mechanism to dynamically select alignment strategies based on the input context.
- Enhance allocentric maps with semantic labels derived from multimodal inputs (e.g., combining textual descriptions with visual data).

Future Potential: Improved alignment and perspective transformation could enhance large-scale reconstructions (e.g., multi-room buildings) and global navigation in robotics.

Local vs. Global Spatial Awareness: **Connection:** While MLLMs excel at local reasoning (e.g., nearby object relationships), integrating this into global awareness mirrors challenges in large-scale 3D reconstruction.

Details:

- Cognitive maps could act as modular components stitched into global 3D representations, solving scalability issues in large environments.
- 3D generation systems could use these local models for finer details while relying on global constraints (e.g., room dimensions).

Transformer^2 Integration:

- Develop modular *expert vectors* for local and global spatial reasoning.
- Dynamically combine these vectors during inference to maintain consistency between local details and global layouts.
- Use SVF's compositional properties to seamlessly integrate multiple cognitive maps into unified representations.

Future Potential: Combining cognitive maps withTransformer^2 could improve realism and scalability in spatial reconstruction.

Dynamic Spaces and Temporal 3D Reasoning: **Connection:** Videos enable MLLMs to observe spaces dynamically, aligning with temporal reasoning in 3D reconstruction or generation.

Details:

- *"Thinking in Space"* involves tracking object positions and predicting changes over time (e.g., a door swinging open).
- Dynamic 3D systems must account for changes to simulate realistic scenarios or create interactive environments.

Transformer^2 Integration:

- Train time-aware *expert vectors* for temporal reasoning (e.g., tracking object movement).
- Use the two-pass inference to dynamically update 3D representations as changes occur.
- Apply RL to optimize performance in scenarios requiring prediction and adaptation to temporal dynamics.
- Explore video frame generation as a Markov-like process where each frame depends on the previous frame, usingTransformer^2's task-specific *expert vectors* to model transitions.

Future Potential: Enable time-aware 3D reconstructions (e.g., showing how a room changes throughout the day) and interactive AR/VR systems with real-time updates.

Multi-modal Data for Enhanced Understanding: **Connection:** MLLMs combine language and vision, while 3D systems rely on multi-modal inputs (e.g., RGB-D, LiDAR, natural language for semantic labeling).

Details:

- MLLMs could bridge gaps in 3D reconstruction by adding semantic layers (e.g., "This is a chair; it should be near the table").
- Multi-modal MLLMs trained on 3D datasets could infer spatial layouts and semantics from diverse inputs.

Transformer^2 Integration:

- Use SVF to fine-tune *expert vectors* on multimodal datasets for tasks like semantic labeling and object recognition.
- Combine text, vision, and spatial data dynamically using Transformer^2's compositional inference capabilities.
- Train *expert vectors* to generate semantically-aware 3D maps (e.g., object-aware reconstructions).

Future Potential: Applications in robotics (e.g., semantic 3D maps) and enhanced object-aware 3D reconstructions.

Research Questions and Transformer^2 Solutions

1. **Egocentric-allocentric alignment:** How can MLLMs reliably transform first-person video data into globally consistent 3D reconstructions?
    - Use Transformer^2's *expert vectors* to dynamically align frames, integrating local and global perspectives for consistent 3D maps.
2. **Unified cognitive map frameworks:** Develop scalable methods for merging local cognitive maps into global spatial models.
    - Leverage Transformer^2's compositionality to stitch modular cognitive maps into unified global models.
3. **Dynamic 3D reconstruction:** Investigate how MLLMs can infer and represent changes in spaces over time for real-time applications.
    - Train Transformer^2 *expert vectors* for temporal reasoning and integrate them into dynamic scene generation pipelines.
4. **Semantic-aware 3D generation:** Combine spatial reasoning with semantic labels to create meaningful 3D spaces (e.g., a "living room" with correctly placed furniture).
    - Use multimodal Transformer^2 vectors to incorporate semantic understanding into 3D reconstructions.
5. **Vision-language-3D fusion:** Explore multi-modal training paradigms where MLLMs integrate visual and textual data to enhance 3D reconstruction/generation.
    - Apply Transformer^2's two-pass mechanism to fuse text and vision inputs for cohesive spatial reasoning.
6. **Temporal 3D reasoning:** Build models that reason about time to reconstruct evolving or interactive environments.
    - Use Transformer^2's RL-optimized *expert vectors* to represent and predict temporal dynamics.
7. **Tools for real-world spaces:** Create tools that adapt virtual overlays to real-world environments using MLLM-based spatial reasoning.
    - Deploy Transformer^2's compositional adaptation for real-time, context-aware overlays.
8. **Spatial AI for autonomous systems:** Develop MLLM-based reasoning for autonomous robots to operate in complex and dynamic real-world spaces.
    - Train Transformer^2 *expert vectors* for navigation, localization, and spatial planning in robotic systems.

9. **VSI-enhanced benchmarks:** Design benchmarks to test MLLMs' ability to reason about global consistency, temporal dynamics, and semantic understanding in 3D environments.
    - Use Transformer^2 to create modular, scalable evaluation pipelines tailored to specific spatial reasoning tasks.
10. **Human-in-the-loop systems:** Research collaborative frameworks where humans and MLLMs jointly refine cognitive maps or 3D reconstructions.
    - Integrate Transformer^2's adaptable *expert vectors* for seamless human-AI interaction in refining spatial models.
11. **Class-Incremental Learning:** Explore how Transformer^2 can handle class-incremental learning by dynamically adapting to new classes without retraining the entire model.
    - Use SVF to isolate and integrate new class-specific representations into the base architecture.

References

[1] Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, Saining Xie, "Thinking in Space: How Multimodal Large Language Models See, Remember, and Recall Spaces", arXiv:2412.14171.

[2] Qi Sun, Edoardo Cetin, Yujin Tang, "Transformer2: Self-adaptive LLMs", arXiv:2501.06252.