# Research Proposal: Visualization-of-Thought with Mixed Modal LLMs

Diabul Haque

January 2026

## 1 Background

### 1.1 Motivation and Theoretical Framework

Chain-of-Thought (CoT) prompting [1] has significantly enhanced the reasoning capabilities of Large Language Models (LLMs) by decomposing complex problems into intermediate textual steps. However, human cognition is not limited to linguistic processing; we use mental imagery to create simulations of the world that allow us to improve our understanding of reality and make better predictions [2]. For example, in his paper introducing the special theory of relativity, Einstein began his argument with a visualization of moving rods, and he has often credited such visualizations with being the birthplace of his theories. [3, 4]

Current Multimodal LLMs (MLLMs) typically treat images and videos as static inputs to be analyzed, rather than dynamic intermediate steps in a reasoning chain. [6] However, with models such as Chameleon, LLMs have the ability to generate interleaved text and images. [7] This opens up the avenue to train these models to perform *Visualization-of-Thought (VoT)*, i.e. CoT integrated with visuals. Taking advantage of this new mixed-modal generation paradigm, Li et. al. were able to fine-tune Chameleon-7B to produce image visualizations of the model's reasoning process. [8]

Soon after Chameleon, we saw a paper titled Transfusion. [9] While Chameleon is purely autoregressive (AR), using an early fusion, discrete token based approach to mixed modal generation, Transfusion uses AR and diffusion for text and images respectively. Building on top of this Transfusion-style approach, Han et. al. created TV2TV - a modeling approach that learns both

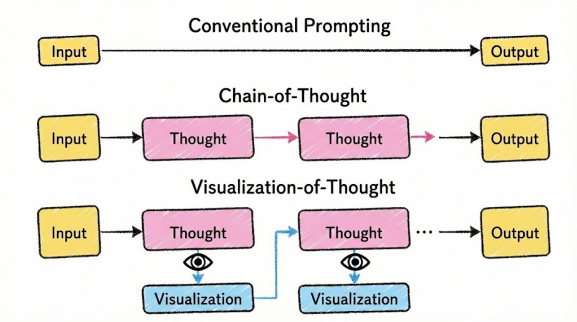


Figure 1: Illustration of Visualization of Thought [5]

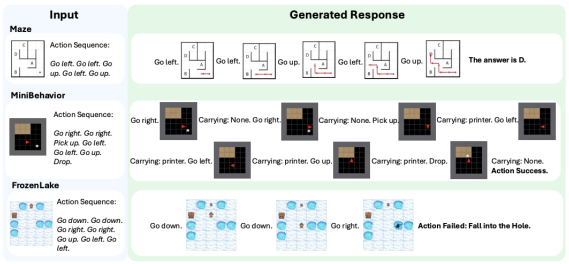


Figure 2: Illustration of Multimodal Reasoning traces [8]

language modeling and video flow matching, and is able to dynamically switch between the two modalities during inference. [10]

While the authors' purpose was to create a model that could *think in words* before producing the next video segment, the paradigm can also work the other way around, i.e. think in videos before producing the desired words. This is important, because humans do not simply think in static images, but are also able to simulate dynamics in our *mind's eye*.

I believe that by leveraging this mixed-modal generation paradigm, and fine-tuning the model to perform VoT using videos, we will able to unlock a new frontier for spatial reasoning and physical understanding.

### 1.2 Research Questions

The primary objective of this MPhil is to develop a mixed modal LLM capable of *thinking* in interleaved text and visuals. Specifically, I intend to investigate:

1. **RQ1:** Can we fine-tune the TV2TV paradigm to generate intermediate visual *thoughts* (VoT), and does this measurably improve performance on spatial reasoning benchmarks compared to text-only CoT?

2. **RQ2:** Can we create diffusion based LLM architectures [11, 12] that can perform interleaved text, image and video generation, and can we fine-tune these architectures to perform VoT?

3. **RQ3:** How do diffusion based LLM architectures compare against AR + Diffusion LLM architectures [10, 13] for interleaved mixed modal generation?

## 2 Methodology

### 2.1 Model Architectures

We will develop and compare two distinct mixed-modal architectures. Both will be initialized using pre-trained weights where possible to ensure strong base capabilities.

#### Model A: The AR + Diffusion Model (TV2TV)

To address **RQ1**, we will adapt the TV2TV framework [10]. This model combines language modeling (next-token prediction) and video flow matching (next-frame prediction) into a single framework using a Mixture-of-Transformers (MoT) architecture [14] with dedicated text and video towers

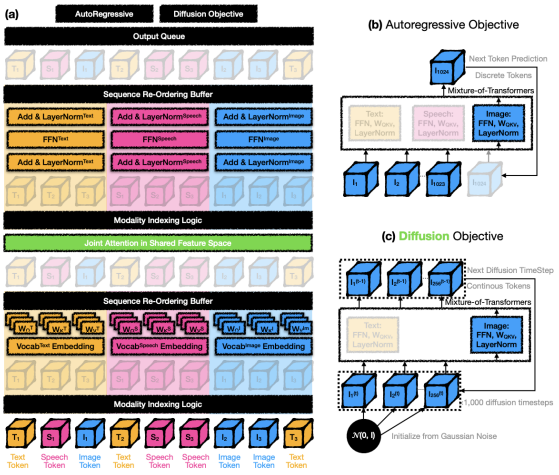


Figure 3: Mixture of Transformers [14]

- **Text Stream:** Standard auto-regressive (AR) next-token prediction.
- **Video Stream:** Flow-matching diffusion for next-frame prediction.
- **Integration:** The model processes interleaved sequences of discrete text and continuous video latents, employing a dual-latent strategy (clean and noisy frames) to enable autoregressive conditioning alongside flow matching.
- **Inference:** During inference, the model dynamically alternates between generating text and video, special tokens (`<BOF>`, `<EOF>`) to switch modalities.

**Model B: The Fully Diffusion Model (Diffusion Text + Diffusion Video)**

To address **RQ2** and **RQ3**, we will modify Model A by replacing the AR text tower with a diffusion-based LLM, similar to Dream-7B or LLaDA [11,15]. I believe this approach will provide an advantage over an unified diffusion based architecture such as MMaDa [12] for the same reasons Transfusion outperforms Chameleon, i.e. because the early fusion models tend to ultimately decouple the processing of the different modalities [14].

## 2.2 Datasets

The datasets that will be utilized in this research need to be structured as interleaved text and video sequences. During the training phase, the model will be trained to generate coherent interleaved text and video. For example, in the TV2TV paper, the authors utilized 8,000 hours of sports data sourced from the YT-Temporal-1B dataset [16]. This dataset was augmented using Vision-Language Models (VLMs) to create video clips interleaved with natural language descriptions. [10]

Per **RQ1**, the goal of our model is to improve our performance on spatial reasoning tasks. To prepare our model for this set of tasks, we will fine-tune it to causal simulations such as collisions, stability tests and dynamics. We will do this by procedurally generating a dataset using physics

engines such as Blender/Unity, augmenting each sample using natural language descriptions from VLMs, and finetuning on the dataset.

## 2.3 Training Pipeline

### Initialization

Each tower will be initialized with weights from open source models of the same modality. For example, we may initialize the AR text tower with Llama-3.1-8B model [17], the diffusion text tower with Dream-7b [11] and the video tower with Wan 2.2 5B [18].

### Supervised Fine-Tuning (SFT)

We will perform fine-tuning on both Model A and Model B using the datasets we curate. We will follow transfusion [9], where the objective of the model will be do minimize losses defined by the standard loss function for each modality used.

### Reasoning Reinforcement Learning

To answer **RQ1**, we must incentivize the model to use video generation not just for aesthetics, but for *utility*. We will employ Group Relative Policy Optimization (GRPO) [19] for Model A which uses AR for text generation and UniGRPO [12] - an algorithm design specifically to incentivize reasoning in diffusion models - for Model B.

- **Policy:** The model generates a video thought $v$ and a text answer $y$.
- **Reward Function ($R$):** We reward **only** the correctness of the final text answer $y$. We do *not* reward the video quality directly.

$$R = I(y = y_{ground\_truth})$$

- **Hypothesis:** By reinforcing only the final outcome, the model will inherently optimize the intermediate video generation to provide the most useful visual cues for solving the problem, effectively "learning to visualize."

## 2.4 Evaluation and Analysis

We will evaluate the models on three dimensions using a hold-out test set and standard benchmarks.

### Spatial and Physical Reasoning (Primary Metric)

We will measure the accuracy difference between Text-Only CoT and Video-VoT on:

- **PHYRE:** A benchmark for physical reasoning in 2D simulations [20].
- **CLEVRER:** Collision Events for Video Representation and Reasoning [21].
- **CausalSpatial:** Anticipating consequences of object motions across four tasks: Collision, Compatibility, Occlusion, and Trajectory [22].

**Efficiency Analysis**

We will compare Model A and Model B regarding the FLOPs and the amount of data necessary for the each model to achieve parity, analyzing the differences in efficiency between the fully diffusion approach and the hybrid AR + Flow matching approach.

# 3 Outcomes and Value

## 3.1 Expected Outcomes

- A comprehensive evaluation framework for "Visualization-of-Thought," establishing whether visual intermediates provide a statistically significant advantage over text-only reasoning.
- A novel Mixed Modal Diffusion LLM architecture capable of state-of-the-art performance on multimodal reasoning benchmarks.
- An open-source codebase and model weights to facilitate further research in multimodal diffusion-based language modeling.

## 3.2 Significance and Value

Models capable of 'visualizing' before acting offer immense value to scientific modeling and robotics. Evolution would not have endowed humans with the ability to visualize if it were not essential for survival. Likewise, an AI's ability to simulate outcomes and reason through them before taking action is may be a crucial requirement for achieving AGI.

## 3.3 Timeline

| **Timeline** | **Milestone** | **Key Deliverable** |
|---|---|---|
| Months 1–2 | Literature & Model Dev | AR+Flow and Dual-Diffusion architectures |
| Months 3–4 | Data Engineering | Curated open source & physics simulation datasets |
| Months 5–7 | SFT | SFT of AR+Flow and Dual-Diffusion architectures |
| Months 8–10 | RL & Optimization | Implementation of GRPO/UniGRPO for reasoning utility. |
| Months 11–12 | Evaluation & Writing | Benchmarking on CausalSpatial/PHYRE and Thesis submission. |

# References


[1] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," 2023.

[2] S. T. Moulton and S. M. Kosslyn, "Imagining predictions: mental imagery as mental emulation," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 364, pp. 1273–1280, May 2009.

[3] A. Einstein, "Zur elektrodynamik bewegter körper," *Annalen der Physik*, vol. 322, no. 10, pp. 891–921, 1905.

[4] Wikipedia, "Einstein's thought experiments — Wikipedia, the free encyclopedia." https://en.wikipedia.org/wiki/Einstein [Online; accessed 25-January-2026].

[5] W. Wu, S. Mao, Y. Zhang, Y. Xia, L. Dong, L. Cui, and F. Wei, "Mind's eye of llms: Visualization-of-thought elicits spatial reasoning in large language models," 2024.

[6] G. Team, "Gemini: A family of highly capable multimodal models," 2025.

[7] C. Team, "Chameleon: Mixed-modal early-fusion foundation models," 2025.

[8] C. Li, W. Wu, H. Zhang, Y. Xia, S. Mao, L. Dong, I. Vulić, and F. Wei, "Imagine while reasoning in space: Multimodal visualization-of-thought," 2025.

[9] C. Zhou, L. Yu, A. Babu, K. Tirumala, M. Yasunaga, L. Shamis, J. Kahn, X. Ma, L. Zettlemoyer, and O. Levy, "Transfusion: Predict the next token and diffuse images with one multimodal model," 2024.

[10] X. Han, Y. Emad, M. Hall, J. Nguyen, K. Padthe, L. Robbins, A. Bar, D. Chen, M. Drozdzal, M. Elbayad, Y. Hu, S.-W. Li, S. D. Roy, J. Verbeek, X. Wang, M. Ghazvininejad, L. Zettlemoyer, and E. Dinan, "Tv2tv: A unified framework for interleaved language and video generation," 2025.

[11] J. Ye, Z. Xie, L. Zheng, J. Gao, Z. Wu, X. Jiang, Z. Li, and L. Kong, "Dream 7b: Diffusion large language models," 2025.

[12] L. Yang, Y. Tian, B. Li, X. Zhang, K. Shen, Y. Tong, and M. Wang, "Mmada: Multimodal large diffusion language models," 2025.

[13] W. Shi, X. Han, C. Zhou, W. Liang, X. V. Lin, L. Zettlemoyer, and L. Yu, "Lmfusion: Adapting pretrained language models for multimodal generation," 2025.

[14] W. Liang, L. Yu, L. Luo, S. Iyer, N. Dong, C. Zhou, G. Ghosh, M. Lewis, W. tau Yih, L. Zettlemoyer, and X. V. Lin, "Mixture-of-transformers: A sparse and scalable architecture for multimodal foundation models," 2025.

[15] S. Nie, F. Zhu, Z. You, X. Zhang, J. Ou, J. Hu, J. Zhou, Y. Lin, J.-R. Wen, and C. Li, "Large language diffusion models," 2025.

[16] R. Zellers, J. Lu, X. Lu, Y. Yu, Y. Zhao, M. Salehi, A. Kusupati, J. Hessel, A. Farhadi, and Y. Choi, "Merlot reserve: Neural script knowledge through vision and language and sound," 2022.

[17] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "Llama: Open and efficient foundation language models," 2023.

[18] T. Wan, "Wan: Open and advanced large-scale video generative models," 2025.

[19] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. K. Li, Y. Wu, and D. Guo, "Deepseekmath: Pushing the limits of mathematical reasoning in open language models," 2024.

[20] A. Bakhtin, L. van der Maaten, J. Johnson, L. Gustafson, and R. Girshick, "Phyre: A new benchmark for physical reasoning," 2019.

[21] K. Yi, C. Gan, Y. Li, P. Kohli, J. Wu, A. Torralba, and J. B. Tenenbaum, "Clevrer: Collision events for video representation and reasoning," 2020.

[22] W. Ma, C. Wang, R. Yuan, H. Chen, N. Dai, S. K. Zhou, Y. Yang, A. Yuille, and J. Chen, "Causalspatial: A benchmark for object-centric causal spatial reasoning," 2026.