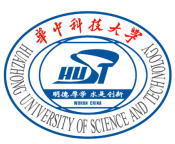
Research Statement

Qingyuan Liu

School of Engineering and Applied Science
Columbia University

Oct. 2025
Homepage: qingyuanliu.net



Outline

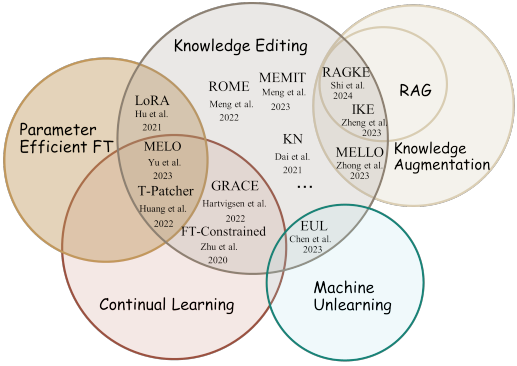
Outline

Title: **Energy-Regularized Sequential Model Editing on Hyperspheres** [arXiv]
In submission: The 14th International Conference on Learning Representations (**ICLR 2026**)
Score (Nov. 23): 8884; Top 2/551 submissions (Transfer/Meta Learning track)

Authors: Qingyuan Liu*, Jia-Chen Gu*, Yunzhi Yao, Hong Wang, Nanyun Peng
Affiliations: PLUSLab, University of California, Los Angeles

*Equal contributions

- Model Editing (Knowledge Editing) aims to refine a pre-trained model by applying one or more edits, where each edit replaces a **factual association** (s, r, o) with new knowledge (s, r, o^*) [1].



Comparison of different technologies

[1] References: [1] Yang, et al. The fall of ROME: understanding the collapse of llms in model editing. EMNLP 2024 [2] Li, et al. Reinforced lifelong editing for language models. CoRR 2025

To achieve this, **locating-and-editing** methods have been proposed for effective model updates. These methods typically follow three steps:

- **Step 1: Locating Influential Layers**: The first step is to identify the specific FFN layers that encode the target knowledge using causal tracing[2].

[2] References: [1] Meng, et al. Locating and editing factual associations in GPT. NeurIPS 2022

To achieve this, locating-and-editing methods have been proposed for effective model updates. These methods typically follow three steps:

- **Step 1: Locating Influential Layers**: The first step is to identify the specific FFN layers that encode the target knowledge using causal tracing[3].

- **Clean run:** with prompt $\boldsymbol{s} + \boldsymbol{r}$.
- **Corrupted run:** obfuscated $\boldsymbol{s}$ randomly.
- **Restoration run:** restore embedding from clean run.

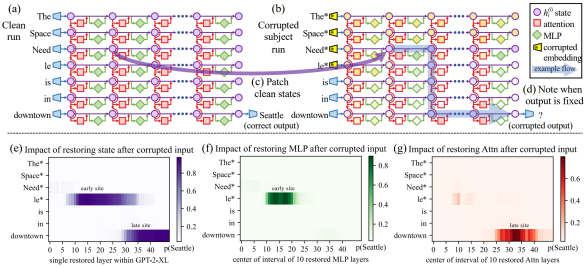
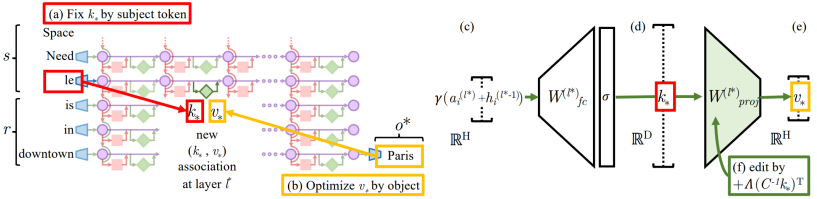


Figure 1: **Causal Traces** compute the causal effect of neuron activations by running the network twice: (a) once normally, and (b) once where we corrupt the subject token and then (c) restore selected internal activations to their clean value. (d) Some sets of activations cause the output to return to the original prediction; the light blue path shows an example of information flow. The causal impact on output probability is mapped for the effect of (e) each hidden state on the prediction, (f) only MLP activations, and (g) only attention activations.

[3] References: [1] Meng, et al. Locating and editing factual associations in GPT. NeurIPS 2022

- **Step 2: Acquiring the Expected Output**: The second step aims to obtain the desired output of the critical layers identified in Step 1.



- Following the **key–value theory**: the key $\boldsymbol{k}$, which encodes $(\boldsymbol{s}, \boldsymbol{r})$, is processed through the output weights$\boldsymbol{W_{out}^l}$ out to produce the original value $\boldsymbol{v}$ encoding $\boldsymbol{o}$.

$$\boldsymbol{V} = \boldsymbol{W}\boldsymbol{K}$$

- **Step 2: Acquiring the Expected Output**: The second step aims to obtain the desired output of the critical layers identified in Step 1.
  - Following the **key–value theory**: the key $\boldsymbol{k}$, which encodes $(\boldsymbol{s}, \boldsymbol{r})$, is processed through the output weights $\boldsymbol{W}_{\boldsymbol{out}}^{\boldsymbol{l}}$ out to produce the original value $\boldsymbol{v}$ encoding $\boldsymbol{o}$.

$$\boldsymbol{V} = \boldsymbol{W}\boldsymbol{K}$$

  - To perform editing, $\boldsymbol{v}$ is expected to be replaced with a new value $\boldsymbol{v}^*$ encoding $\boldsymbol{o}^*$. To this end, current methods typically **use gradient descent on ΔW, maximizing the probability that the model outputs the word associated with o^*** [4].

$$\boldsymbol{V}' = (\boldsymbol{W} + \boldsymbol{\Delta W})\boldsymbol{K}$$

[4] References:[1] Meng, et al. Mass-editing memory in a transformer. ICLR 2023

- **Step 3: Updating W_{out}^l**: This step aims to update the parameters W_{out}^l. It includes a factual set $\{\boldsymbol{K}_1, \boldsymbol{V}_1\}$ containing $\boldsymbol{u}$ new associations, while **preserving the set** $\{\boldsymbol{K}_0, \boldsymbol{V}_0\}$ containing $\boldsymbol{n}$ original associations. Specifically,

$$\begin{aligned} \boldsymbol{K}_0 &= [\,\boldsymbol{k}_1\ \boldsymbol{k}_2\ \cdots\ \boldsymbol{k}_n\,], \quad \boldsymbol{V}_0 = [\,\boldsymbol{v}_1\ \boldsymbol{v}_2\ \cdots\ \boldsymbol{v}_n\,], \\ \boldsymbol{K}_1 &= [\,\boldsymbol{k}_{n+1}\ \boldsymbol{k}_{n+2}\ \cdots\ \boldsymbol{k}_{n+u}\,], \quad \boldsymbol{V}_1 = [\,\boldsymbol{v}_{n+1}^*\ \boldsymbol{v}_{n+2}^*\ \cdots\ \boldsymbol{v}_{n+u}^*\,] \end{aligned} \tag{1}$$

- **Step 3: Updating W_{out}^l**: This step aims to update the parameters W_{out}^l. It includes a factual set $\{\boldsymbol{K}_1, \boldsymbol{V}_1\}$ containing $\boldsymbol{u}$ new associations, while **preserving the set** $\{\boldsymbol{K}_0, \boldsymbol{V}_0\}$ containing $\boldsymbol{n}$ original associations. Specifically,

$$\begin{aligned} \boldsymbol{K}_0 &= [\,\boldsymbol{k}_1\ \boldsymbol{k}_2\ \cdots\ \boldsymbol{k}_n\,], \quad \boldsymbol{V}_0 = [\,\boldsymbol{v}_1\ \boldsymbol{v}_2\ \cdots\ \boldsymbol{v}_n\,], \\ \boldsymbol{K}_1 &= [\,\boldsymbol{k}_{n+1}\ \boldsymbol{k}_{n+2}\ \cdots\ \boldsymbol{k}_{n+u}\,], \quad \boldsymbol{V}_1 = [\,\boldsymbol{v}_{n+1}^*\ \boldsymbol{v}_{n+2}^*\ \cdots\ \boldsymbol{v}_{n+u}^*\,] \end{aligned} \tag{2}$$

Based on these, the objective can be defined as: [5]

$$\tilde{\boldsymbol{W}}_{\boldsymbol{out}}^l \triangleq \arg\min_{\hat{\boldsymbol{W}}} \left(\sum_{i=1}^{n} \left\| \hat{\boldsymbol{W}} \boldsymbol{k}_i - \boldsymbol{v}_i \right\|^2 + \sum_{i=n+1}^{n+u} \left\| \hat{\boldsymbol{W}} \boldsymbol{k}_i - \boldsymbol{v}_i^* \right\|^2 \right). \tag{3}$$

[5] References:[1] Meng, et al. Mass-editing memory in a transformer. ICLR 2023

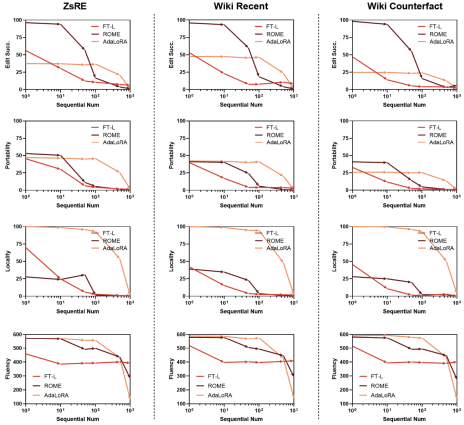
Challenges in Model Editing

★ **Lifelong Editing** (large-scale sequential edit): How to maintain the efficacy of edit while preserving the **general ability** of the edited model?

- **Multi-hops Editing** (reasoning editing): How to maintain the editing efficacy in related **multi-hop questions**?
- **Non-structural Knowledge Editing**: How to generalize current method in general knowledge format **except for factual association** (s, r, o).
- **Editing on Emerging Architectures**: How can model editing be effectively applied to **new architectures** (e.g., multimodal, MoE, or sparse-activated models) where knowledge is distributed across diverse modules and modalities?

Challenges in Model Editing

★ **Lifelong Editing** (large-scale sequential edit): How to maintain the efficacy of edit while preserving the **general ability** of the edited model?[6]



Sequential editing results in randomly selected data from WikiData$_{counterfact}$, ZsRE -and WikiData$_{recent}$ with different numbers.

[6] References: [1] Zhang, et al. A Comprehensive Study of Knowledge Editing for Large Language Models. preprint 2023

★ **Lifelong Editing** (large-scale sequential edit): How to maintain the efficacy of edit while preserving the **general ability** of the edited model?[7]

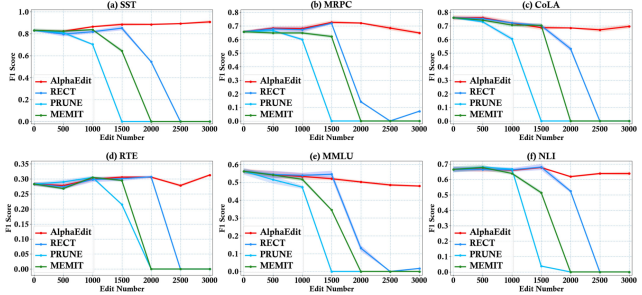


Figure 4: F1 scores of the post-edited LLaMA3 (8B) on six tasks (*i.e.*, SST, MRPC, CoLA, RTE, MMLU and NLI) used for general capability testing. Best viewed in color.

[7] References: [1] Fang, et al. ALPHAEDIT: NULL-SPACE CONSTRAINED KNOWLEDGE EDITING FOR LANGUAGE MODELS. ICLR 2025

From Angular Perspective

- Network generalization by alleviating redundancy through **angular diversification**.[8]

$$\min_{\mathcal{W}} \mathcal{L}(\mathcal{W}) \quad \text{s.t.} \quad 1 \leq i < j \leq m, \quad \left| \frac{\boldsymbol{w}_i \cdot \boldsymbol{w}_j}{\|\boldsymbol{w}_i\|_2 \, \|\boldsymbol{w}_j\|_2} \right| \leq \tau$$

- Why Does **Orthogonal Transformation** Make Sense?

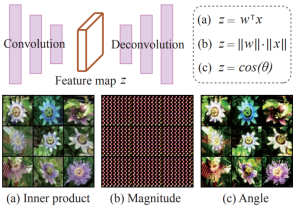


Figure 2: A toy experiment to demonstrate the importance of angular information. The autoencoder is trained in a standard way using inner product activation, and (a) shows the standard reconstruction. In testing, the angular information of neurons alone can well recover the input image, even if the autoencoder is not trained with angles.

[8] References:[1] Xie, D. et al. Learning latent space models with angular constraints. In ICML, 2017. [2] Cogswell, Ahmed. et al. Reducing overfitting in deep networks by decorrelating representations. In ICLR, 2016. [3] Qiu, L. et al. Controlling Text-to-Image Diffusion by Orthogonal Finetuning. NeurIPS, 2023.

Research Questions

- 1. Does **knowledge editing** affect the **angular diversity** of edited weights?
- 2. Is there any correlation between **angular diversity** and {**editing, general task**} performance?
- 3. If **Yes**, is it possible to design an **angular diversity**-based regularization method to improve knowledge editing?

Before Introducing Correlation...

Mathematical Definition[9]:

- **Hyperspherical Energy (HE)** is a metric that quantifies the **uniformity of vector distributions on a hypersphere**.
- Given a normalized metric $\boldsymbol{X}$ (to only focus on angular aspect), the $\boldsymbol{HE}$ could be calculated by:

$$\boldsymbol{HE}(\boldsymbol{X}) = \sum_{i \neq j} \left(\|\hat{\boldsymbol{x}}_i - \hat{\boldsymbol{x}}_j\|^2 + \epsilon \right)^{-s} = \sum_{i \neq j} \left(2(1 - \cos\theta_{ij}) + \epsilon \right)^{-s}$$

[9] References:[1] Liu, L. et al. Learning Towards Minimum Hyperspherical Energy. NeurIPS, 2020.

Before Introducing Correlation...

Mathematical Definition:

- **Hyperspherical Energy (HE)** quantifies the **uniformity of vector distributions on a hypersphere**.
- Given a normalized metric $\mathbf{X}$ (focusing on angular aspect), the **HE** can be calculated by:

$$\mathrm{HE}(\mathbf{X}) = \sum_{i \neq j} \left(\|\hat{\mathbf{x}}_i - \hat{\mathbf{x}}_j\|^2 + \epsilon \right)^{-s} = \sum_{i \neq j} \left(2\left(1 - \cos\theta_{ij} \right) + \epsilon \right)^{-s}$$

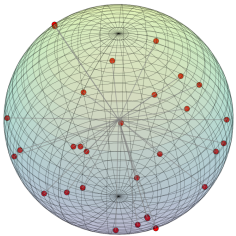
When vectors are close to orthogonal (around 90° apart), their cosine similarity diminishes toward 0, so this term approaches 1 and the overall **HE** decreases.

- **High HSE** $\rightarrow$ Vectors are **clustered** and non-uniformly distributed.
- **Low HSE** $\rightarrow$ Vectors are **well-dispersed** and uniformly distributed.

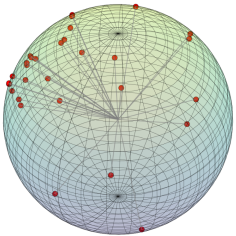
Before Introducing Correlation...

Mathematical Definition:

- **Hyperspherical Energy (HE)** quantifies the **uniformity of vector distributions on a hypersphere**.



(a) Weight Neurons (low HE)



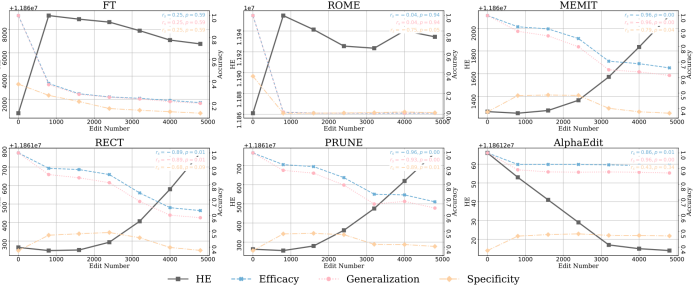
(b) Weight Neurons (high HE)

- **High HSE** $\rightarrow$ Vectors are **clustered** and non-uniformly distributed.
- **Low HSE** $\rightarrow$ Vectors are **well-dispersed** and uniformly distributed.

Research Questions

- 2. Does **knowledge editing** affect the **angular diversity** of edited weights?
- 3. Is there any correlation between **angular diversity** and {**editing, general task**} performance?

- **Observation 1:** **Collapse in sequential editing is closely tied to sharp fluctuations in HE.**

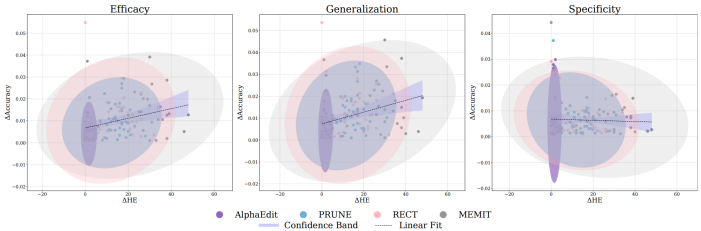


- **5,000** sequential edits with batch size of **100** on **LLaMA3-8B**.

Figure 2: Trends of HE and editing performance throughout sequential editing. The Spearman correlation scores between HE and each editing metric displayed at the end of each curve.

From Angular Perspective

- **Observation 2:** **Advanced editing methods suppress HE fluctuations effectively.**



- The correlation between **changes** in **HE (HE)** and editing performance **(Acc.)**, where **each point denotes the difference between two consecutive batch edits**.

Figure 3: Correlation between changes in HE and editing performance across consecutive edited weights. Each point corresponds to a ΔHE–ΔAcc. pair for one method over five thousand sequential edits. Confidence ellipses and regression lines illustrate overall trends.

From Angular Perspective

- **Theoretically:** $|\Delta \mathrm{HE}|$ **is the lower-bound constraint for** $\Delta \mathbf{V}$ **(disruption caused by editing).**

$$
\begin{aligned}
For\ \Delta \mathbf{V} &= (\mathbf{W} + \Delta \mathbf{W})\mathbf{K} - \mathbf{W}\mathbf{K} \\
&= \mathbf{W}\mathbf{K} + \Delta \mathbf{W}\mathbf{K} - \mathbf{W}\mathbf{K} \\
&= \mathbf{V} + \Delta \mathbf{V} - \mathbf{V} \\
&= \Delta \mathbf{V}
\end{aligned}
$$

which is the variation on the original knowledge ($\Delta \mathbf{V}$)

From Angular Perspective

- **Theoretically: $|\Delta \mathrm{HE}|$ is the lower-bound constraint for $\Delta \mathbf{V}$ (disruption caused by editing).**

$$
\begin{aligned}
For\ \Delta \mathbf{V} &= (\mathbf{W} + \Delta \mathbf{W})\mathbf{K} - \mathbf{W}\mathbf{K} \\
&= \mathbf{W}\mathbf{K} + \Delta \mathbf{W}\mathbf{K} - \mathbf{W}\mathbf{K} \\
&= \mathbf{V} + \Delta \mathbf{V} - \mathbf{V} \\
&= \Delta \mathbf{V}
\end{aligned}
$$

Theorem 1 (Lower Bound on Output Perturbation).[10] *Under the assumptions of orthonormal inputs and small perturbations, the output perturbation $\Delta \mathbf{V}$ is lower-bounded by squared change in HE:*

$$
|\Delta \mathbf{V}| \geq \left(\frac{\Delta \mathrm{HE}}{K} \right)^2, \qquad K = 4 \left(\sum_{k=1}^{p} \left(\sum_{j \neq k} \|\mathbf{w}_k - \mathbf{w}_j\|^{-3} \right)^2 \right)^{1/2}
$$

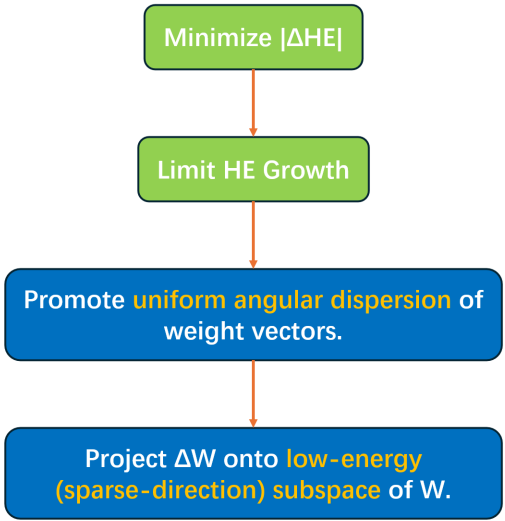
where K is a constant dependent on the original weight matrix geometry.

[10] [1] NOTE: Detailed proof in our paper Section 3/Appendix in ENERGY-REGULARIZED SEQUENTIAL MODEL EDITING ON HYPERSPHERES: https://arxiv.org/abs/2510.01172

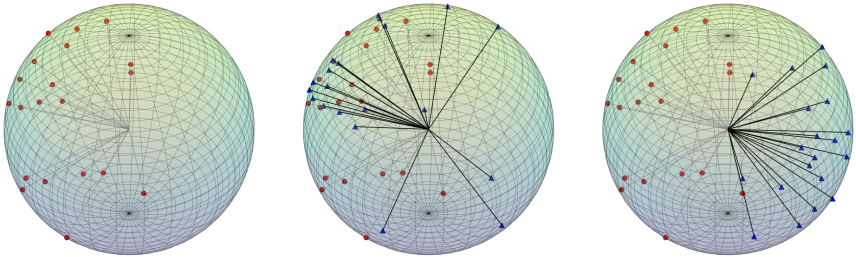
Research Questions

- 3. If **Yes**, is it possible to design an **angular diversity**-based regularization method to improve knowledge editing?

From Angular Perspective



- We introduce **SPHERE** (**S**parse **P**rojection for **H**yperspherical **E**nergy **R**egularized **E**diting)



(a) Weight Neurons (b) Current Methods (c) SPHERE (Ours)

Figure 1: (a) A weight matrix is viewed as a set of neurons (red dots) on a hypersphere. (b) Current SOTA methods (Ma et al., 2025; Fang et al., 2025) introduce perturbations (blue triangles) that interfere with the principle hyperspherical directions of pre-edit weights. (c) SPHERE projects new knowledge onto a sparse space complementary to the principal hyperspherical directions.

- **1. Principal Space Estimation:** Estimate the principal space of $\mathbf{W}$ via Rayleigh Quotient optimization.

$$\mathbf{v} = \arg\max_{\|\hat{\mathbf{v}}\|=1} \left(\frac{1}{n}\|\mathbf{W}\hat{\mathbf{v}}\|^2\right) = \arg\max_{\|\hat{\mathbf{v}}\|=1} \left(\frac{1}{n}\hat{\mathbf{v}}^\top(\mathbf{W}^\top\mathbf{W})\hat{\mathbf{v}}\right)$$

- **1. Principal Space Estimation:** Estimate the principal space of $\mathbf{W}$ via Rayleigh Quotient optimization.

$$\mathbf{v} = \arg\max_{\|\hat{\mathbf{v}}\|=1} \left(\frac{1}{n}\|\mathbf{W}\hat{\mathbf{v}}\|^2\right) = \arg\max_{\|\hat{\mathbf{v}}\|=1} \left(\frac{1}{n}\hat{\mathbf{v}}^\top(\mathbf{W}^\top\mathbf{W})\hat{\mathbf{v}}\right)$$

According to the Rayleigh Quotient theory:

Key Insight

The maximum value of $Proj(\mathbf{v})$ is the largest eigenvalue of $\frac{1}{n}\mathbf{A}^\top\mathbf{A}$, and the corresponding direction $\mathbf{v}$ is the **principal eigenvector**.

Then we can construct the principal subspace matrix based on the top-r eigenvectors of $\frac{1}{n}\mathbf{A}^\top\mathbf{A}$:

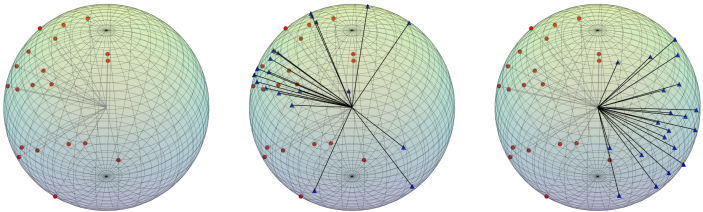
$$\mathbf{U} = [\,\mathbf{v}_{d-r+1}, \ldots, \mathbf{v}_d\,] \in \mathbb{R}^{d\times r}$$

- **2. Sparse Space Definition:**

$$P_{\perp} = I - \alpha \mathbf{U}\mathbf{U}^{\top} \in \mathbb{R}^{d \times d}$$

- **3. Sparse Space Definition:**[11]

$$\hat{\mathbf{W}} = \mathbf{W} + \Delta \mathbf{W}_{\text{proj}} = \mathbf{W} + \Delta \mathbf{W} P_{\perp}$$



(a) Weight Neurons (b) Current Methods (c) SPHERE (Ours)

[11] References: [1] we also provide a mathematical proof that SPHERE suppresses the HE, ensuring bounded variations in the hidden representations $\Delta \boldsymbol{V}$: see Appendix C.2 in our paper: https://arxiv.org/abs/2510.01172

Research Questions

- **RQ1:** How does SPHERE perform on sequential editing **compared to baseline** methods?
- **RQ2:** Can SPHERE effectively **preserve the hyperspherical uniformity** of edited weights?
- **RQ3:** How does SPHERE-edited LLMs perform on **general ability** evaluations?
- **RQ4:** Can baseline methods be significantly improved with **plug-and-play SPHERE**?

Analysis of SPHERE

- **RQ1:** How does SPHERE perform on sequential editing tasks **compared to baseline** methods?

Table 1: Comparison of SPHERE with existing methods on sequential editing. *Eff.*, *Gen.*, *Spe.*, *Flu.* and *Consis.* denote Efficacy, Generalization, Specificity, Fluency and Consistency, respectively. The best results are highlighted in bold, while the second-best results are underlined.

Method	Model	ZSRE			Counterfact				
		Eff.↑	Gen.↑	Spe.↑	Eff.↑	Gen.↑	Spe.↑	Flu.↑	Consis.↑
Pre-edited		35.42±0.30	34.17±0.30	38.02±0.27	0.49±0.07	0.44±0.05	18.09±0.24	634.84±0.12	22.06±0.08
FT	LLaMA3-8B	15.27±0.21	14.78±0.21	5.06±0.10	8.40±0.28	2.54±0.13	0.03±0.01	409.80±0.67	19.35±0.13
MEMIT	LLaMA3-8B	0.00±0.00	0.00±0.00	0.06±0.01	0.00±0.00	0.00±0.00	0.00±0.00	318.19±0.24	4.19±0.04
PRUNE	LLaMA3-8B	10.35±0.18	10.08±0.18	9.55±0.15	1.19±0.11	0.34±0.04	0.62±0.03	**618.72±0.08**	**49.24±0.13**
RECT	LLaMA3-8B	0.01±0.00	0.01±0.01	0.04±0.01	0.57±0.08	0.29±0.04	0.10±0.01	438.83±0.18	9.40±0.05
AlphaEdit	LLaMA3-8B	86.64±0.23	81.28±0.28	28.78±0.22	4.37±0.20	1.71±0.10	0.57±0.03	482.36±0.44	4.71±0.04
SPHERE	LLaMA3-8B	**90.01±0.21**	**84.67±0.26**	**45.40±0.29**	**52.89±0.50**	**32.07±0.39**	**5.01±0.10**	551.51±0.53	30.89±0.13
Pre-edited		35.29±0.29	34.10±0.28	38.44±0.27	0.42±0.06	0.46±0.05	15.06±0.20	624.45±0.11	23.02±0.69
FT	Qwen2.5-7B	4.97±0.14	4.58±0.13	4.01±0.11	15.44±0.36	4.63±0.17	1.46±0.05	214.26±0.09	3.15±0.02
MEMIT	Qwen2.5-7B	0.13±0.02	0.12±0.01	0.04±0.01	0.00±0.00	0.00±0.00	0.00±0.00	370.84±0.30	3.59±0.03
PRUNE	Qwen2.5-7B	47.93±0.36	45.50±0.35	**39.20±0.28**	14.30±0.35	11.27±0.26	**6.75±0.12**	**620.74±0.10**	**29.50±0.08**
RECT	Qwen2.5-7B	0.73±0.04	0.75±0.04	0.05±0.07	0.64±0.08	0.19±0.03	0.09±0.01	368.46±0.27	1.35±0.01
AlphaEdit	Qwen2.5-7B	42.01±0.40	39.99±0.39	13.87±0.20	43.92±0.50	24.37±0.36	2.32±0.06	479.83±0.77	4.67±0.07
SPHERE	Qwen2.5-7B	**70.04±0.36**	**65.43±0.37**	27.35±0.26	**60.76±0.49**	**29.24±0.37**	3.83±0.08	612.67±0.22	14.74±0.07

- **15,000** edits on **LLaMA3 (8B)**, **5,000** edits on **Qwen2.5 (7B)**.
- SPHERE achieves substantial gains in both **Efficacy** and **Generalization**, with average improvements of **24.19%** and **16.02%**, respectively, over the best baseline.

- **RQ2:** Can SPHERE effectively **preserve the hyperspherical uniformity** of edited weights?

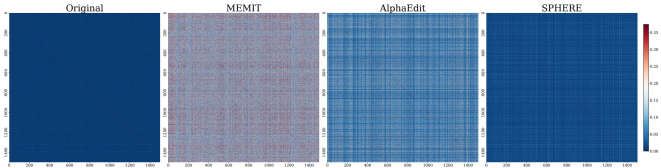


Figure 4: Cosine similarity between neurons in the updated weight matrix after 15,000 edits. Darker colors indicate lower similarity, reflecting better hyperspherical and orthogonal uniformity. SPHERE effectively preserves the weight structure, demonstrating the most stable hyperspherical uniformity.

- SPHERE effectively preserves hyperspherical uniformity after editing, as the **cosine similarity** among weight neurons remains **close to the original distribution**, thereby avoiding directional collapse.

- **RQ2:** Can SPHERE effectively **preserve the hyperspherical uniformity** of edited weights?

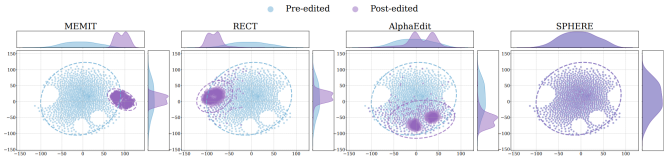


Figure 5: The t-SNE distribution of weight neurons of pre-edited and post-edited LLM after 15,000 edits using dimensionality reduction. The top and right curve graphs display the marginal distributions for two reduced dimensions, where SPHERE consistently exhibits minimal shift.

- The pre- and post-edited weights exhibit **nearly overlapping distributions**, indicating that SPHERE prevents significant shifts in weights and maintains consistency.

- **RQ3:** How does SPHERE-edited LLMs perform on **general ability** evaluations?

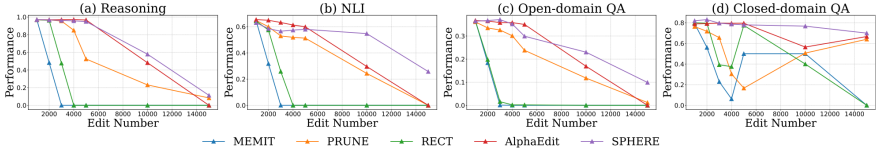
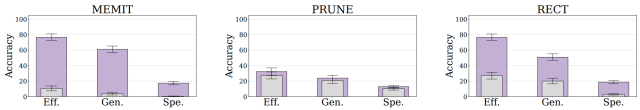


Figure 6: General ability testing of post-edited LLaMA3 (8B) on four tasks.

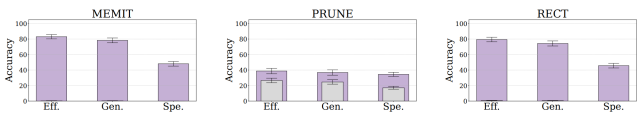
- **Four representative tasks** were adopted following RECT[12], including **Reasoning** on GSM8K, **Natural Language Inference (NLI)** on RTE, **Open-domain QA** on Natural Questions, and **Closed-domain QA** on BoolQ.
- SPHERE effectively **preserves the general abilities** of post-edited LLMs even under extensive editing, maintaining the original model performance across all metrics **after 15k edits.**

[12] References:[1] Model Editing Harms General Abilities of Large Language Models: Regularization to the Rescue

- **RQ4:** Can baseline methods be significantly improved with **plug-and-play SPHERE**?



(a) Performance Improvement on Counterfact Dataset



(b) Performance Improvement on ZsRE Dataset

Figure 7: Performance improvements of baseline editing methods after adding a single line of code from SPHERE (i.e., sparse space projection). Gray bars denote the original baseline performance, while purple bars indicate the performance after enhancement.

- On average, the optimized baselines achieve relative improvements of **49.05%**, **42.64%**, and **24.44%** in **Efficacy**, **Generalization**, and **Specificity**, respectively **after adding a single line of code from SPHERE (i.e., sparse space projection)**.

- **RQ4:** Can baseline methods be significantly improved with **plug-and-play SPHERE**?

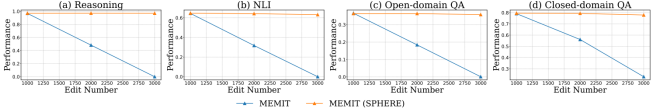


Figure 10: General ability improvements of MEMIT after incorporating SPHERE with a single line of sparse space projection code.)

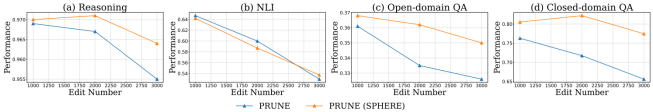
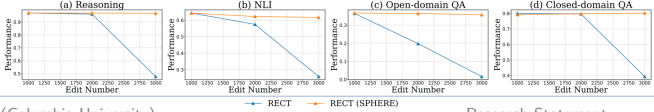


Figure 11: General ability improvements of PRUNE after incorporating SPHERE with a single line of sparse space projection code.)



- The baselines enhanced with the SPHERE projection also demonstrate **significantly better robustness in general abilities**.

Outline

Outline

Title: **Causal Adjacency Learning for Spatiotemporal Prediction Over Graphs** [arXiv]
Accepted: 27th IEEE International Conference on Intelligent Transportation Systems (**ITSC 2024**)
Authors: Zhaobin Mo*, **Qingyuan Liu***, Baohua Yan, Longxiang Zhang, and Xuan Di
Affiliations: DitecT Lab, Columbia University

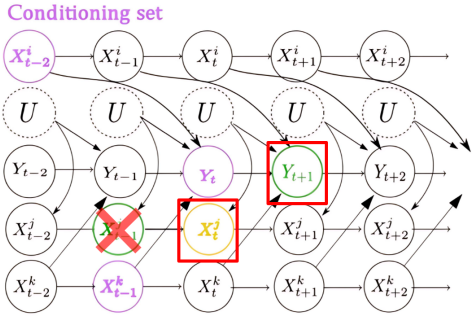
*Equal contributions

- Most studies calculate the **adjacency matrix** by directly memorizing the data, such as **distance-** and **correlation**-based matrices.
- These adjacency matrices ignore potential distribution shifts between training and test data, leading to the **Out-of-Distribution (OOD)** generalization problem.

Research Question

Can we find the underlying causal relations in temporal dimension to enhance the prediction on learned graph?

- **Kernel-based Conditional Independence Test (CIT)** —identifies dependencies among **temporal variables** via kernel similarity measures.
- **Systematic Path Isolation (SyPI)**[13]: filters out **spurious or fake relations** discovered by the kernel-based CIT.



Check if the following conditions hold true

$$X^j_l \not\perp\!\!\!\perp Y_{t+1} | X^i_{t-2}, X^k_{t-1}, Y^t \quad (1)$$

$$X^j_{t-1} \perp\!\!\!\perp Y_{t+1} | X^j_l, X^i_{t-2}, X^k_{t-1}, Y^t \quad (2)$$

$\not\perp\!\!\!\perp$: dependent

$\perp\!\!\!\perp$: independent

[13] References: [1] Necessary and sufficient conditions for causal feature selection in time series with latent common causes, ICML 2021

Causal Adjacency Learning

- **Key idea:** **Replace the traditional adjacency matrix** with a **learned causal adjacency matrix** to address the **Out-of-Distribution (OOD)** issue.

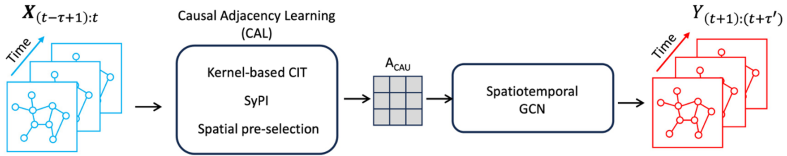
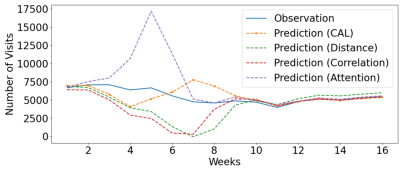


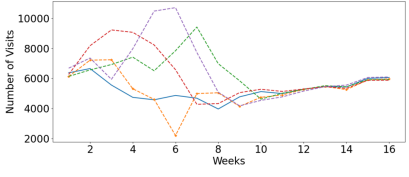
Fig. 1: Framework of the upstream CAL and the downstream spatiotemporal GCN for the problem of STPG.

Experiment

Compared with our **distance-**, **correlation-**, and **attention-** based baseline matrix, the CAL algorithm has a better performance in predicting the OOD dataset, particularly for **long-term predictions**.



(a) Predict next 1 week (ZIP = 10309)



(b) Predict next 4 week (ZIP = 10309)

Experiment

- **CAL method** achieves **14.23%**, **15.49%**, and **50.27%** RMSE improvement in predicting the **next week's mobility** in New York, compared with **distance-**, **correlation-**, and **attention-** based adjacency matrices, respectively.

TABLE II: Performance metrics of GCN prediction on future 1-4 weeks based on different adjacency matrices. Bolded represents the best result and underlined means second best.

#Adjacency	RMSE/MAE				
	T+1	T+2	T+3	T+4	Avg.
Distance	.221/.049	.284/.081	.339/.115	.414/.171	.322/.104
Correlation	.224/.050	.258/.066	.259/.067	.337/.114	.273/.074
Attention	.381/.145	.276/.076	**.233/.054**	.412/.170	.334/.111
CAL(Ours)	**.189/.036**	**.239/.057**	.239/.057	**.292/.085**	**.243/.059**

Outline

Title: **Balanced Latent Space of Diffusion Models for Counterfactual Generation** [arXiv]
Accepted: The Thirteenth International Conference on Learning Representations Deep Generative Model in Machine Learning: Theory, Principle and Efficacy Workshop (**ICLR DeLTa 2025**)

Authors: Baohua Yan, **Qingyuan Liu**, Zhaobin Mo, Kangrui Ruan, Xuan Di
Affiliations: DitecT Lab, Columbia University

COLUMBIA UNIVERSITY
IN TIMINE TVO VIDEBIMVS LVMEN

*Equal contributions

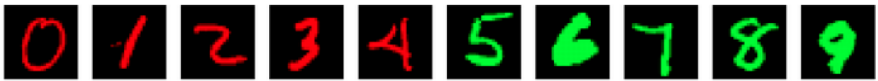
Motivation

- **Goal:** To address the **Out-of-Distribution (OOD)** problems in vision models.
- In the **MNIST_Colored** dataset, the training set contains digits **0–4** in red and **5–9** in green, while the test set reverses this color mapping (**0–4** in green, **5–9** in red).

MNIST_Colored Dataset (Training set)

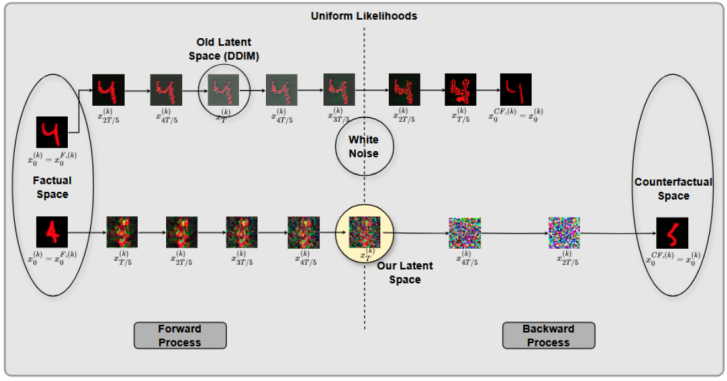
Motivation

- **Goal:** To address the **Out-of-Distribution (OOD)** problems in vision models.
- In the **MNIST_Colored** dataset, the training set contains digits **0–4** in red and **5–9** in green, while the test set reverses this color mapping (**0–4** in green, **5–9** in red).



MNIST_Colored Dataset (Training set)

- Due to the OOD issue, it is **challenging to use supervised models effectively.** To address this, we aim to generate **counterfactual data** e.g., a red **"5"** or a green **"0"** to augment the training set and improve generalization.



Balanced Latent Space. A latent point x_T is equally close to its factual and counterfactual counterparts:

$$d(x_T, x_0^F) = d(x_T, x_0^{CF}).$$

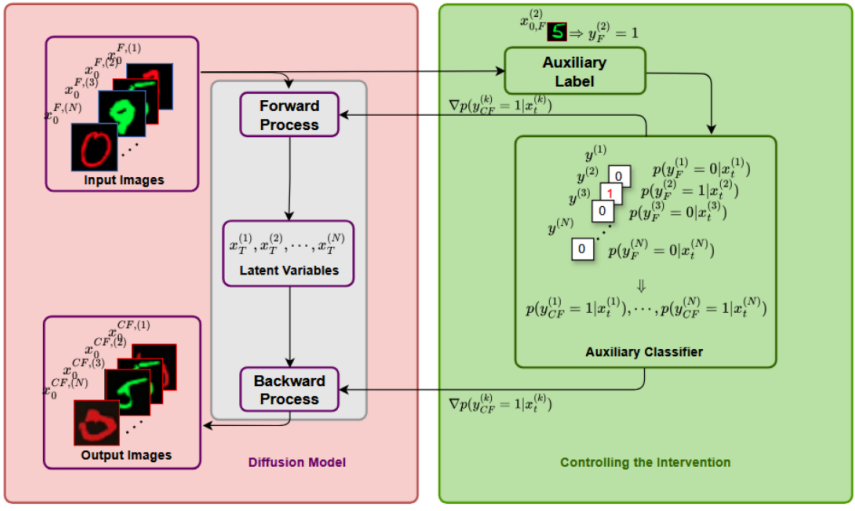
Forward Process Toward Balanced Latent Space

Update Rule:

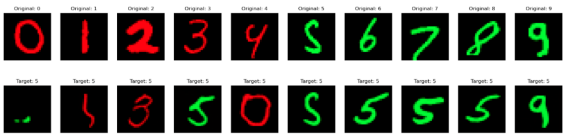
$$\Delta x_t = \varepsilon_\theta(x_t + \zeta_t \nabla_{x_t} p_\phi(y_{CF} \mid x_t),\, t) - \varepsilon_\theta(x_t, t)$$

$$x_{t+1} = x_t + \gamma_1 \Delta x_t + \gamma_2 z, \qquad z \sim \mathcal{N}(0, I)$$

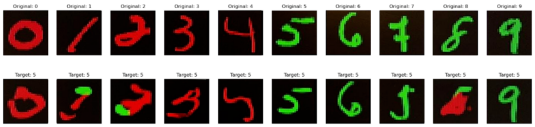
Intuition:
Guided diffusion step that pushes x_t toward the region where $p(y_F \mid x) = p(y_{CF} \mid x)$ (balanced latent space).


Input Images
Forward Process
Latent Variables
Backward Process
Output Images
Diffusion Model
Auxiliary Label
Auxiliary Classifier
Controlling the Intervention

Experiment & Ablation Study



Generation results with old latent space



Generation results with new latent space. Note that the red "5" is **strongly OOD**.

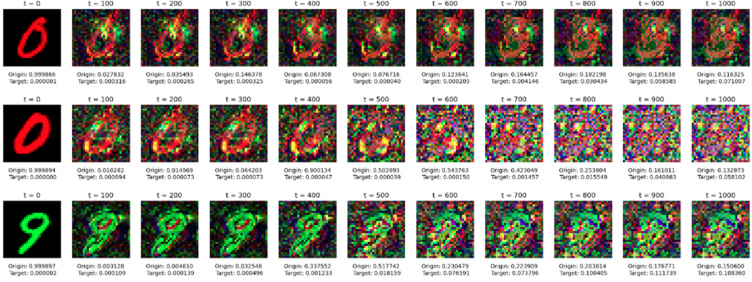


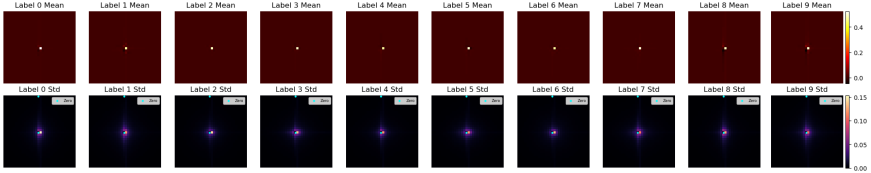
Figure 2: Sample likelihoods at different time during the forward process. We randomly select 3 different digits and compare their likelihoods of being original digit and target digit.

Outline

Title: **InSPECT: Invariant Spectral Features Preservation of Diffusion Models**
In submission: to **CVPR 2026 Authors**: Baohua Yan, **Qingyuan Liu**, Jennifer Kava, Xuan Di
Affiliations: DitecT Lab, Columbia University

*Equal contributions

- We find that there exists a frequency component **invariant** across all classes (CIFAR10 dataset).



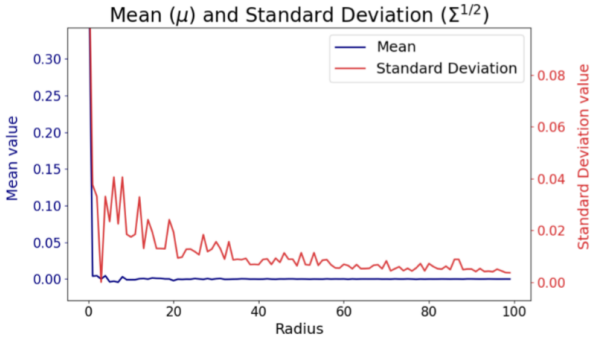


Figure 5. Mean and standard deviation of the CIFAR-10 dataset. x-axis is the frequency radius of the spectral coefficients. We pick red channel and show the curves related to one label.

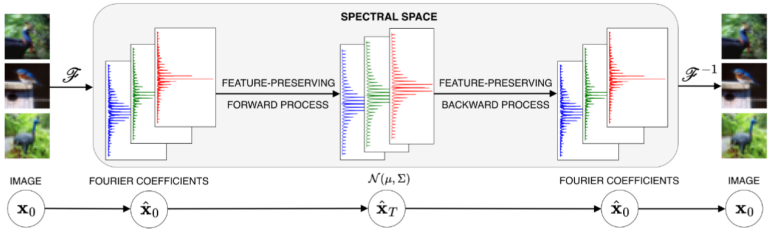


Figure 4. **The graphical model of Invariant Spectral Feature Preservation in Diffusion Models (InSPECT) considered in this work.** We convert a given image dataset, $\mathbf{x}_0$ into a Fourier coefficients and carry out the InSPECT forward and backward process, and convert the Fourier coefficients back into $\mathbf{x}_0$.

- $\epsilon \sim \mathcal{N}(0, \mathbf{I}), \qquad q(x_t \mid x_0) = \mathcal{N}\big(\sqrt{\bar{\alpha}_t} x_0,\ (1 - \bar{\alpha}_t)\mathbf{I}\big)$
- $\epsilon_{\text{dir}} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \qquad q(\hat{x}_t \mid \hat{x}_0) = \mathcal{N}\big(\sqrt{\bar{\alpha}_t} \hat{x}_0 + (1 - \sqrt{\bar{\alpha}_t})\boldsymbol{\mu},\ (1 - \bar{\alpha}_t)\boldsymbol{\Sigma}\big)$

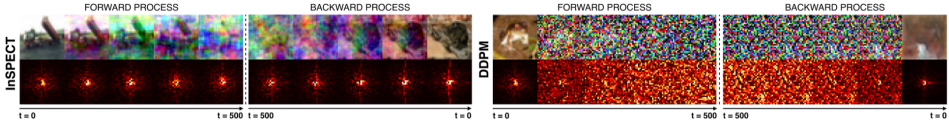
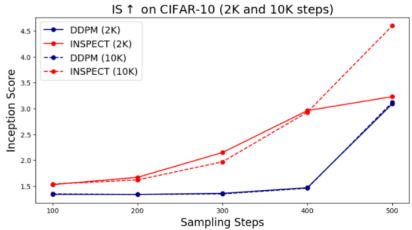
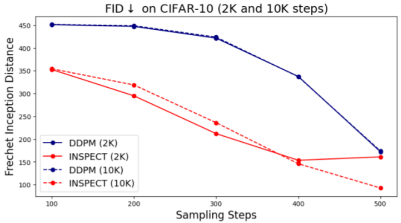


Figure 6. Visual comparison of forward and backward processes between InSPECT (**left**) and DDPM (**right**) by illustration of the images (**top row**) with their corresponding Fourier coefficients (**bottom row**) at different diffusion time-steps. InSPECT guides the image toward a specified random noise instead of a white noise, demonstrating significantly better generative quality.



(a) Inception Score (IS ↑) for different sampling steps under 2K and 10K training iterations.



(b) Fréchet Inception Distance (FID ↓) for different sampling steps under 2K and 10K training iterations.

Outline

Outline

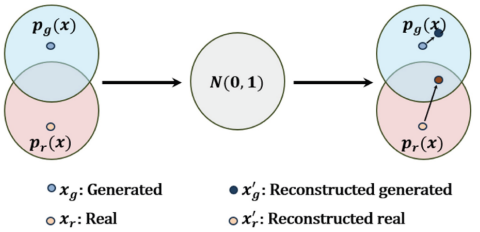
Title: **Turns Out I'm Not Real: Towards Robust Detection of AI-Generated Videos**
[arXiv]
Highlight: **Columbia Engineering Research Highlight**
Accepted: In IEEE / CVF Computer Vision and Pattern Recognition Conference 2024, GenAI workshop. (**CVPR GenAI 2024**)
Authors: **Qingyuan Liu**, Pengyuan Shi, Yun-Yun Tsai, Chengzhi Mao, and Junfeng Yang
Affiliations: Software Systems Lab, Columbia University

*Equal contributions

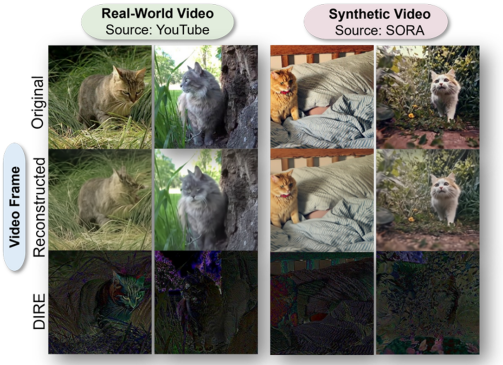
What is DIRE? And what is DIREs role in AI-Synthetic Detection?

- **Diffusion Reconstruction Error (DIRE)** helps distinguish between **human-generated** and **diffusion-generated** content.
- It measures the **difference between the input frame and its reconstruction** from a diffusion model:

$$DIRE(x_0) = |x_0 - \mathbf{R}(\mathbf{I}(x_0))|$$



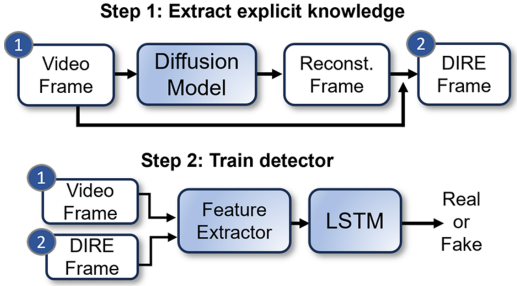
- **Real** video frames are sampled from **YouTube**, while **synthetic (fake)** videos come from **SORA** by OpenAI.
- The **reconstructed** frame of a **SORA-generated** video is visually **closer** to the input frame, whereas the **real** video from **YouTube** are not (e.g., distorted cat face).



Dataset

	Video Source	Denoising Condition	Generated Model	# of Clips (real/fake)
In-domain	VidVRD [18]	Image2Video	SVD-XT [9]	2k/2k
Out-domain	VidVRD (test)	Image2Video	Pika [2]	107/107
	VidVRD (test)	Image2Video	Gen-2 [3]	107/107
	YouTube/SORA [6]	x	x	207/191

- **The Flow of DIVID**



- Given a sequence of video frames, we first **generate the reconstructed version** of each frame using a diffusion model.
- **Compute the DIRE** (Diffusion Reconstruction Error) between each reconstructed frame and its input, revealing the reconstruction discrepancy.
- A **CNN+LSTM** detector jointly models RGB frames and DIRE sequences to classify videos as **Real or Synthetic**.

Experiment & Ablation Study

	Detector Architecture	Evaluation Metrics	
		Acc.	AP
RGB	CNN	90.16	97.02
RGB	CNN+LSTM	90.16	97.39
DIRE [22]	CNN	92.74	97.46
DIVID/ DIRE only	CNN+LSTM	**93.68**	97.66
DIVID/ DIRE + RGB	CNN+LSTM	91.33	**98.20**

	Model	Out-domain			Total Avg.
		Gen-2	Pika	SORA	
RGB	CNN	65.42	78.04	60.05	67.84
RGB	CNN+LSTM	**67.76**	84.11	60.80	70.89
DIRE [22]	CNN	50.93	60.75	54.77	55.48
DIVID / DIRE only	CNN+LSTM	60.75	80.37	60.8	67.3
DIVID / DIRE + RGB	CNN+LSTM	66.82	**86.92**	**61.01**	**71.58**

- DIVID achieves **98.20%** average precision (AP) and has better detection accuracy, and outperforms them by **0.94%** to **3.52%**.
- On three **out-domain** test sets, including **SORA**, **Pika**, and **Gen-2**. DIVID improves the out-domain average accuracy by **0.69%** to **16.1%**.

Outline

Title: **LAVID: An Agentic LVLM Framework for Diffusion-Generated Video Detection** [arXiv]

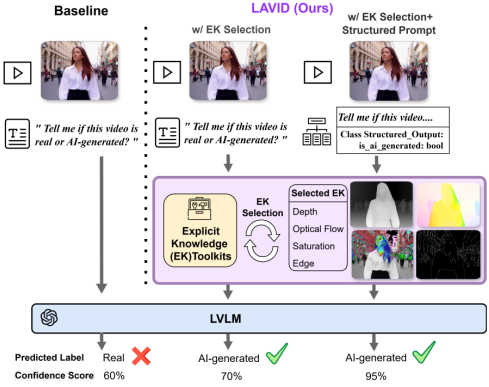
Status: Preprint (***ICCV 2025*** *score: 442*)

Authors: **Qingyuan Liu**, Yun-Yun Tsai, Ruijian Zha, Pengyuan Shi, Victoria Li, Chengzhi Mao and Junfeng Yang

Affiliations: Software Systems Lab, Columbia University

*Equal contributions

- **Deep Learning** based detection frameworks always faced with limitations like **transparency**, **inability to recognize new artifacts**



- We proposed an **self-evolving LVLM Framework** for Diffusion-Generated Video Detection with explicit knowledge enhancement.

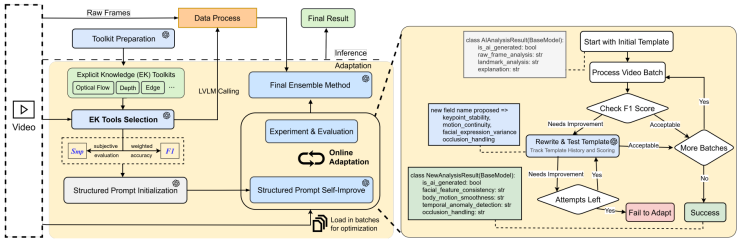


Figure 2. An agentic framework (**LAVID**) for video detection. The left part shows our main pipeline. First, LVLMs suggest tools relevant to video detection, and based on the model's preferences and the performance improvement each tool provides, we assemble a customized toolkit for each LVLM for video detection. The right part shows the details of the online adaptation for structured prompt. The prompt tuning will be based on the LVLM itself. Component marked with the logo are developed with the LVLM like GPT-4o [41].

- We create a new benchmark called VidForensic which features 200 text-to-video prompts and more than **1.4k** high-quality videos, collected or generated from **eight generative models**.

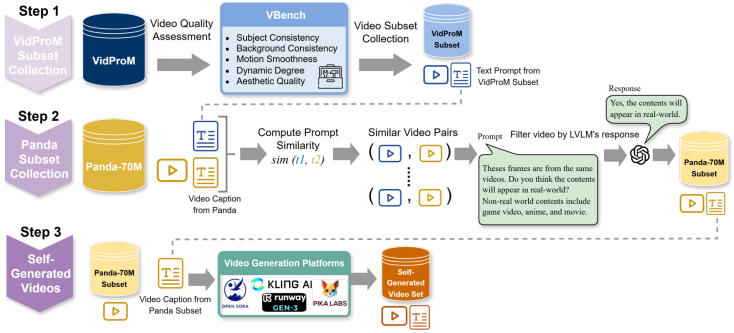


Figure 6. Dataset collection pipeline for VidForensic. Component marked with the logo are developed with the LVLM like GPT-4o [41].

Experiment

- Evaluation results show that LAVID improves F1 scores by **6.2 to 30.2%** over the top baselines on our benchmark across four SOTA LVLMs, including Llava, Qwen-VL, Gemini-1.5-pro, GPT-4o.

LVLM	Method	VidForensic (VidProM) [52]				VidForensic (Self-collected)				Avg.
		Pika [1]	T2vz [33]	Vc2 [18]	Ms [49]	OpenSORA [59]	Gen3 [2]	Kling [3]	SORA [11]	
Llava-OV-7B [36]	Baseline1 (w/o SP)	53.50/14.68	61.00/37.10	61.00/37.10	58.50/30.25	52.50/12.11	50.00/1.96	50.00/1.96	54.44/16.33	55.12/18.94
	Baseline2 (w/o SP)	50.50/1.98	51.00/3.92	51.50/5.83	53.50/13.08	52.00/7.69	50.00/0.00	50.00/0.00	50.00/0.00	51.06/4.06
	Baseline3 (w/o SP)	**54.50/18.02**	62.00/39.68	65.00/46.97	62.00/39.68	54.00/16.36	51.00/5.77	50.00/1.96	55.56/20.00	56.76/23.56
	LAVID (w/o SP)	**54.50/18.02**	**70.00/57.75**	**69.00/55.71**	**68.00/53.62**	**58.00/28.81**	**51.50/7.62**	**50.50/3.88**	**55.56/20.00**	**59.63/32.69**
Qwen-VL-Max [44]	Baseline1 (w/o SP)	72.50/63.09	75.00/67.53	82.00/78.57	76.00/69.23	67.50/53.24	62.00/40.62	54.50/19.47	58.89/39.34	68.55/51.24
	Baseline2 (w/o SP)	60.50/38.76	75.00/68.35	71.50/62.25	72.50/64.05	60.50/38.76	52.00/14.29	50.00/7.41	56.67/26.42	62.33/39.56
	Baseline3 (w/o SP)	74.00/67.90	79.00/75.58	84.50/83.06	**79.50**/76.30	69.50/60.13	65.50/52.41	54.00/24.59	61.11/47.76	70.89/60.97
	LAVID (w/o SP)	**87.00/88.39**	**81.50/82.63**	**86.00/87.39**	77.00/**77.45**	**79.00/79.81**	**82.50/83.72**	**60.00/52.94**	**67.78/71.84**	**77.60/76.08**
Gemini-1.5-pro [22]	Baseline1 (w/o SP)	68.33/54.32	71.00/59.72	67.00/51.47	75.00/67.11	68.50/54.68	64.00/44.62	58.00/28.81	58.89/41.27	66.34/49.83
	Baseline2 (w/o SP)	73.50/66.24	81.00/77.91	76.00/70.37	85.00/83.33	71.50/62.75	71.50/62.75	59.50/37.21	71.11/64.86	72.51/58.28
	Baseline3 (w/o SP)	64.50/45.80	77.00/70.51	71.00/59.72	76.50/69.68	64.50/45.80	62.00/39.68	52.50/11.21	61.11/42.62	66.08/51.28
	LAVID (w/o SP)	**92.00/91.73**	**96.33/96.38**	**95.83/95.87**	**97.50/97.56**	**92.17/91.93**	**88.50/87.67**	**74.83/68.46**	**76.67/78.36**	**89.23/88.43**

LVLM	Method	VidForensic (VidProM) [52]				VidForensic (Self-collected)				Avg.
		Pika [1]	T2vz [33]	Vc2 [18]	Ms [49]	OpenSORA [59]	Gen3 [2]	Kling [3]	SORA [11]	
GPT-4o [41]	Baseline1 (w/ SP)	89.00/89.22	90.00/90.29	**92.50**/92.89	85.00/84.69	82.50/81.68	86.00/85.86	66.50/57.86	68.89/64.10	82.55/80.82
	Baseline2 (w/ SP)	72.00/77.95	70.00/76.00	71.00/76.98	66.50/72.43	68.00/73.98	68.00/73.98	64.50/70.29	65.56/70.84	68.20/74.06
	Baseline3 (w/ SP)	89.50/88.66	90.50/90.73	92.00/92.31	86.00/85.71	82.00/80.85	85.00/84.54	69.00/61.73	63.33/50.75	82.17/79.41
	LAVID (w/ SP)	**93.00/93.46**	91.50/91.94	**92.50**/92.96	89.00/89.32	**86.50**/86.57	**91.00**/91.43	75.50/72.63	68.89/68.89	85.99/85.90
	LAVID (OA w/ SP)	91.50/92.17	**92.00/92.52**	**92.50/93.02**	**90.50/91.24**	**86.50/86.79**	**91.00/91.59**	**77.00/76.77**	**70.93/72.11**	**86.49/87.03**

- Comparison with supervised learning methods and Application on deepfake.

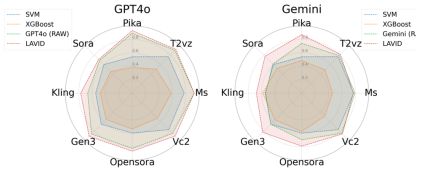


Figure 4. Comparison between supervised learning methods and LAVID. Both SVM and XGBoost are trained with the same EK of the LVLMs. (RAW) represents the results using raw frame only.

Method	Trainset	Celeb-DF-v1	
		Acc.	F1
Guo et al. [24]	FF++ [46]	73.19	–
RECCE [12]	FF++ [46]	71.81	–
MAT [58]	FF++ [46]	71.81	–
Baseline (Gemini-1.5-pro)	–	44.00	17.65
Baseline (GPT-4o)	–	64.95	74.24
LAVID (Gemini-1.5-pro) w/ Face-Seg	–	50.00	37.50
LAVID (GPT-4o) w/ Face-Seg	–	**75.00**	**80.91**

Table 6. Performance comparison of existing Deepfake detection baselines, the baseline prompts, and LAVID on Celeb-DF-v1. Video-level accuracy (Acc.) and F1-score (F1) are used as evaluation metrics where available. The reported performance of RECCE and MAT are referenced from [51].

Outline

Outline

Title: **Unified Vision-Language Foundation Model for Brain MRI Interpretation**
On Going Project: The research is expected to be submitted to **Nature**
Affiliations: VLAA LAB, University of California, Santa Cruz

*Equal contributions

Motivation

- Developed large-scale vision foundation models for 3D brain imaging (i.e., T1, MPRAGE, DWI), trained on 800+ public datasets with **170,000** scans (billion-level slices);

Dataset Name	Anatomy	Modality	Adult/Ped	Disease
BaptistHealth_R01Intake	Brain	Radiology	Adult	Glioma
BraTS 2023: Segmentation - Adult Glioma (training a	Brain	Radiology	Adult	Glioma
BURDENKO	Brain	Radiology	Adult	Glioma
CCF	Brain	Radiology	Adult	Glioma
IVYGAP	Brain	Radiology	Adult	Glioma
CPTAC-GBM	Brain	Radiology	Adult	Glioma
LUMIERE	Brain	Radiology	Adult	Glioma
Preprocessed_Brigham	Brain	Radiology	Adult	Glioma
TCIA TCGA-GBM	Brain	Radiology	Adult	Glioma
UCSF-PDGM	Brain	Radiology	Adult	Glioma
UCSF-POST	Brain	Radiology	Adult	Glioma
xCures	Brain	Radiology	Adult	Glioma
ACRIN-DSC-MR-Brain	Brain	Radiology	Adult	Glioma
ACRIN-FMISO-Brain	Brain	Radiology	Adult	Glioma
GLIS-RT	Brain	Radiology	Adult	Glioma
Qin GBM Treatment Response	Brain	Radiology	Adult	Glioma
IU_PrimaryBrainTumor	Brain	Radiology	Adult	Glioma
FETS GBM Cohort	Brain	Radiology	Adult	Glioma
REMBRANDT	Brain	Radiology	Adult	Glioma
Brain-Tumor-Progression_TCIA	Brain	Radiology	Adult	Glioma
GBM_Cohort (Brucegroup)	Brain	Radiology	Adult	Glioma
Erasmus Glioma Database	Brain	Radiology	Adult	Glioma
Preprocessed_MCW-adult_marwa	Brain	Radiology	Adult	Glioma
CPTAC-GBM	Brain	Pathology	Adult	Glioblastoma
TCGA-GBM	Brain	Pathology	Adult	Glioblastoma
IVYGAP	Brain	Pathology	Adult	Glioblastoma
BraTS-Path Challenge 2024	Brain	Pathology	Adult	Glioblastoma

Dataset Name	Anatomy	Modality	Adult/Ped	Disease	IDiA Steward	Download Source
ADRC	Brain	Radiology	Adult	AD	Apoorva	UW
WRAP	Brain	Radiology	Adult	AD	Apoorva	UW
IRAP	Brain	Radiology	Adult	AD	Apoorva	Sheba Medical Center
UKBiobank	Brain	Radiology	Adult	AD	Apoorva	UK Biobank

Dataset Name	Contributing Hospitals	Access Means
Baptist-BrainMets	Miami cancer center	DUA
Brain-TR-GammaKnife		TCIA
CCF		DUA
HenryFordHealth		DUA
nnUNet_raw_data_base	r trial train nnunet. Please ignore.	Publicly Available
NYUmets		DUA
Ocana-Tienda et al Nature Paper Data		Publicly Available
MOTUM	ty, Changzheng Hospital, and The First Affiliated Hospital of USTC	Publicly Available
UCSF-BMSR		TCIA
UH		DUA

- **Vision Tasks**

Category	Task	Public benchmark	Metric
Brain-age regression	Estimate “brain age” from 3D MRI	UK Biobank MRI test split; Cam-CAN; IXI; ABCD; PAC2019 challenge	MAE
Diagnosis / classification	- HGG / LGG / Normal- Glioma / Meningioma / Metastasis- AD / MCI / HC- Brain-age regression	BraTS-Class, TCIA-Meningioma, ADNI + OASIS, UK Biobank	ROC-AUC, MAE
Lesion segmentation	Tumor sub-regions, WMH, MS lesions, stroke core, ICH, aneurysm	BraTS-Seg, WMH 2017, MSSEG, ISLES 2022, RSNA-ICH, ADAM	Dice
Tissue segmentation	GM/WM/CSF, skull-stripping	MRBrainS18, SkullStrip-12k	Dice
Registration	Replace DINOv2 features in pairwise deformable reg	LPBA40, CANDI	TRE, DSC
Reconstruction	T1 → FLAIR synthesis, 2× super-res, motion-deblur	Calabrese, FastMRI-Brain	PSNR / SSIM
Anomaly detection	One-class tumor / hemorrhage	BraTS-OOD, CQ500-ICH	AUROC

- **Language Tasks**

Mode	Task idea	Dataset / how to build	Metric
CLIP-style retrieval	Image ↔ report contrastive	Build from RadReports-MRI, NYU MRNet notes	Recall@K
VQA (3-D)	Yes/No & span answers (“Is there midline shift?”)	Convert reports to Q-A pairs; evaluate on VQARad-Brain	Acc
Caption / report generation	Autocomplete “Findings” section	Same corpus, ROUGE-L, BLEU	
Text-conditioned segmentation	Prompt: “segment enhancing core”	BraTS masks + textual cues	Dice
Zero-shot classification	Natural-language labels (“metastatic lesion”)	Use held-out classes from BraTS, ADNI	AUROC

Thank You For Your Listening

Qingyuan Liu
ql2505@columbia.edu